Analysis of DNA-Transposons in the Genome of *Brachypodium distachyon* Reveals Mechanisms for Intergenic Sequence Turnover

---

Dissertation

zur
Erlangung der naturwissenschaftlichen Doktorwürde
(Dr. sc. nat.)

vorgelegt der

Mathematisch-naturwissenschaftlichen Fakultät

der Universität Zürich

von

Jan Piotr Buchmann

von

Zürich

Promotionskomitee

Prof. Dr. Beat Keller (Vorsitz)
PD Dr. Thomas Wicker (Leitung der Dissertation)
Prof. Dr. Christian von Mering

Zürich 2012

"L'absurde n'a de sens que dans la mesure ou l'on n'y consent pas"
Albert Camus, *Le Mythe de Sisyphe*, 1942

"If I have seen further it is by standing on ye sholders of Giants"
Isaac Newton, *Letter to Robert Hooke*, 1676

"Look behind you, a three-headed monkey!"
Guybrush Threepwood, *The Secret of Monkey Island*, 1990

Contents

List of Figures

List of Tables

Summary

Amongst the recently sequenced plant genomes is *Brachypodium distachyon*, after rice and sorghum the third completely sequenced grass genome. It was the first sequencing project where TEs were annotated by a special consortium, the *Brachypodium* Repeat Annotation Consortium (BRAC), for which we annotated the *CACTA* DNA-Transposons. We identified and characterized 14 *CACTA* families which contributed approximately 3% of the *B. distachyon* genome. The *CACTA* data have been expanded with additional data from *Zea mays*, *Oryza sativa*, *Triticum aestivum*, *Arabidopsis thaliana* and *Petunia hybrida*. Phylogenetic analysis and comparison of conserved intron/exon boundaries between different *CACTA* transposases resulted in an evolutionary model where ancient *CACTA* transposases were exon rich and loss of introns is a major evolutionary mechanism.

We were also interested in mechanisms which drive the turnover in intergenic sequences. Sequence turnover is described as balance between amplification of DNA through TE activity and DNA loss through unequal crossing over (UECO) and illegitimate recombination (IR). The molecular mechanisms of UECO are well studied. In contrast, the molecular basis of IR was largely unknown. To investigate intergenic sequence turnover, we compared five orthologous loci between *Brachypodium distachyon* and *Brachypodium sylvaticum* spanning 1 mega base pair (Mbp) in total. We estimated the divergence time between *B. distachyon* and *B. sylvaticum* to be approximately 1.7–2.0 million year ago (MYA). While the majority of genes was found in colinear positions, the intergenic space has undergone a virtually complete turnover. DNA-Transposon excision sites in one or the other species

revealed highly diagnostic double-strand break (DSB) repair motifs. Depending on the DSB repair mechanism, excision of DNA-Transposons deleted hundreds of base pair (bp) of flanking sequence (Single Strand Annealing) or inserted several hundred bp of "filler DNA" (Synthesis Dependent Strand Annealing). In some cases, DSBs were repaired by a combination of both. In total, the identified events exchanged 17% of the intergenic space. We developed a model for the evolution of the intergenic space where the repair of the DSB upon DNA-Transposon excision is a major force, explaining most of the observed IR signatures.

Additionally, we refined methods to estimate divergence times between two species based on CDS, intergenic and LTR-retrotransposon sequences. Until now the type of sequence used for divergence time estimations, e.g. intergenic or CDS, required the selection of the according substitution rate. Our method is based on the removal of sites in sequence alignments which have accelerated or decreased substitution rates in codons or intergenic sequences, allowing the use of only one substitution rate.

Zusammenfassung

Das kürzlich publizierte Genom von *Brachypodium distachyon* ist das dritte komplett sequenzierte Genom in der Familie der Gräser. Gleichzeitig war es das erste Sequenzierprojekt mit einem eigenem Konsortium für die Annotation der repetitiven Elemente (RE), dem *Brachypodium* Repeat Annotation Consortium (BRAC). Unsere Aufgabe in diesem Konsortium war die Annotation der *CACTA* DNA Transposons. Wir identifizierten und charakterisierten 14 *CACTA* Familien welche ungefähr 3% des ganzen Genoms ausmachten. Der Datensatz wurde mit *CACTA* Elementen aus *Zea mays*, *Oryza sativa*, *Triticum aestivum*, *Arabidopsis thaliana* und *Petunia hybrida* erweitert. Dies ermöglichte eine Analyse der konservierten Exon/Intron Grenzen sowie die phylogenetische Rekonstruktion der Transposasen. Das Resultat ist ein Modell zur Evolution der *CACTA* Transposasen, das von einer intronreichen Ur-Transposase ausgeht, die im Laufe der Zeit Introns verliert.

Ebenfalls untersuchten wir die Mechanismen welche zu der schnellen Divergenz von intergenischen Regionen führen. Der Austausch der intergenischen Regionen ist als Gleichgewicht zwischen der Amplifikation von DNA durch RE und Deletionen durch ungleiches Crossing Over (UECO) und illegitime Rekombination (IR) beschrieben. Der molekulare Mechanismus von UECO ist bekannt, im Gegensatz zur IR mit bis jetzt unbekannter molekularer Basis. Den Vergleich der intergenischen Regionen führten wir auf fünf orthologen Loci mit total 1 Megabasenpaar zwischen *B. distachyon* and *Brachypodium sylvaticum* durch. Wir datierten die Divergenzzeit zwischen den zwei Arten, welche ungefähr 1.7–2.0 Millionen Jahre beträgt (MYA). Im Gegensatz zu den Genen, welche mehrheitlich in kollinearen Positionen gefunden

wurden, waren die intergenischen Regionen nahezu komplett verschieden. Positionen, an welchen DNA Transposon ausgeschnitten wurden, zeigten spezifische Doppelstrangbruch (DSB) Reparaturmotive. Je nach DSB Reparaturmechanismus hat das Ausschneiden von DNA Transposons hunderte von benachbarten Basenpaaren (bp) gelöscht (Single Strand Annealing) oder eingefügt (Synthesis dependent Strand Annealing). In manchen Fällen wurden DSBs durch eine Kombination der beiden Mechanismen repariert. Das Ausschneiden von DNA Transposons und die darauf folgende Reparatur des DSB haben 17% der intergenischen Regionen ausgetauscht. Basierend auf der Aktivität von DNA Transposons entwickelten wir ein Modell zur Erkärung des schnellen Austauschs der intergenischen Regionen. Unser Modell kann auch die bis jetzt unbekannte Enstehung von IR Signaturen erklären.

Zusätzlich haben wir Methoden zur Datierung von Divergenzzeiten mit Hilfe von CDS (protein codierenden), intergenischen sowie LTR-Retrotransposon Sequenzen optimiert. Bis jetzt war die Wahl der Substitutionsrate zur Datierung der Divergenzzeit von der verwendeten Art der Sequenzen abhängig. Unsere Methode basiert auf dem Entfernen von Nukleotiden in Sequenzvergleichen mit beschleunigter oder herabgesetzter Substitutionsrate. Dies erlaubt die Verwendung von nur einer Substitutionsrate, unabhängig von den verwendeten Sequenzen.

General Introduction *

1

The study of genome evolution on the molecular level is tightly linked to available DNA sequences of an organism. In an ideal case, this is a complete genomic sequence of an organism or a dataset of partial sequences in good quality, i.e. long sequences with a high coverage. Much of higher plant genomes consists of transposable elements (TEs) or other repetitive elements. Thus, the length and quality of genomic sequences is crucial in analyzing the role and influence of TEs in the evolution of genomes or, in this thesis, grass genomes. Therefore, it is not surprising that this kind of analysis goes hand in hand with the progress in whole genome analysis tools and sequencing techniques.

1–1 Plant Genomes and Genome Sequencing

The first completely sequenced plant genome was from *Arabidopsis thaliana*, a dicotyledonous model plant with a genome size of approximately 120 mega base pairs (Mbp) (AGI, 2000). This genome was sequenced by the "BAC-by-BAC" approach, a technique which consists in constructing a bacterial artificial chromosome (BAC) library of the genome, fingerprinting the BAC clones and assembling them into a minimum tiling path which is then sequenced by shotgun-sequencing. While the result is a high quality sequence which is ordered along the chromosomes, it is very laborious and expensive. This method was also used to sequence the genomes of *Sorghum bicolor* (Paterson *et al.*, 2009), soybean (*Glycine max*, Schmutz *et al.* 2010) and *Zea mays* (Maize Schnable *et al.* 2009). Maize, with a genome size of

*Parts of this introduction has been published in a book chapter by Buchmann *et al.* (in press)

2,300 Mbp is the largest plant genome sequenced so far (Schnable *et al.*, 2009).

Only two years after the published *A. thaliana* genome sequence, the complete sequences of the first grass (monocotyledon) genomes were published: *Oryza sativa* L ssp. *japonica* and *Oryza sativa* L ssp. *indica* (Goff *et al.*, 2002; Yu *et al.*, 2002). These projects used a whole-genome shotgun (WGS) approach where the genome is randomly broken up into small sequence segments which are sequenced and the resulting reads are assembled. This method became especially feasible due to the development in DNA sequencing technology which resulted in faster and cheaper methods.

The introduction of next generation sequencing technologies (NGS) around 2008 led to a significant reduction in the price per sequenced base pair and whole genome (Wetterstrand, 2012). This allowed *de novo* genome sequencing and assembly of e.g. the genomes of cacao (Argout *et al.*, 2011), *Brassica rapa* (Wang *et al.*, 2011) and *Brachypodium distachyon* (IBI, 2010). Interestingly, the number of bases in GenBank from 1982 until today has doubled approximately every 18 months (Benson *et al.*, 2008). This rate is similar to Moore's law which states that the number of transistors which can be placed on a integrated circuit doubles approximately every 18 to 24 months (Moore, 1965). Still, large and complex plant genomes, e.g. barley or wheat, push WGS to its limits due to the repetitive nature of those genomes. In such cases, anchoring the shotgun sequences or individual BACs to genetic maps is an essential procedure.

The goal of each sequencing project is to retrieve so called "pseudomolecules" for each chromosome from the sequenced organism. These pseudomolecules are not complete strings of bases from telomere to telomere, but contain gaps where highly repetitive regions such as the centromere, ribosomal DNA clusters or TEs interfere with the sequencing and/or assembly process. This nuisance is observed in all applied sequencing techniques until now. To close those gaps, genomic

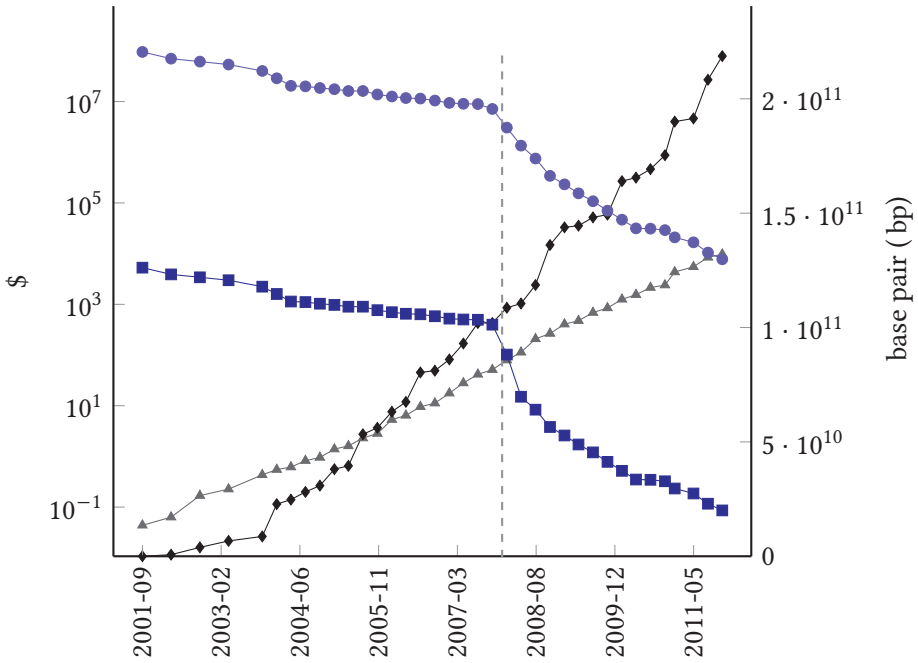


Figure 1.1 Comparison of sequencing costs and sequence deposition in GenBank since September 2001.
The X-axis depicts the time range between September 2001 and May 2011. The left Y-axis depicts the costs per sequenced Mbp (—■—) or sequenced genome (—●—). Note the logarithmic scale for the Y-axis. The right X-axis depicts the number of bps stored in GenBank, either in the main database (—▲—) or in the separate Whole Genome Sequence database (—♦—). The dashed line indicates the introduction of NGS mthods, e.g. 454, Solexa. Sources: Wetterstrand (2012); Benson *et al.* (2008).

sequences are under continuous improvement after the initial release. The high quality genomes from rice and *Arabidopsis thaliana* which have reached the seventh and tenth version, respectively, still contain gaps (Table 1.1). The bottleneck in the newer sequencing techniques like 454 and Illumina, which create extensive sequence information in every run, is the final assembly which remains still a challenge for existing software.

The faster and cheaper techniques boosted genome sequencing. At the time of writing there were 25 plant genome projects listed on `http://www.phytozome.org`, of which five are from the grass family (Goodstein *et al.*, 2012). Not all published genomes have the same high quality standards. The genomes of poplar, *Vitis vinifera* (grapevine) and *Physcomitrella patens* are based on only one round of shotgun sequencing, resulting in unordered "supercontigs" or "scaffolds" with thousand of sequence gaps and little anchoring to chromosomes. Chain *et al.* (2009) suggested a classification system with five categories, reflecting the used sequencing technology and quality of the assembly. The categories range from 1 to 5, where 1 represents the lowest Standard Draft (basic automated assembly of raw sequences) and 5 the highest category for genome sequences (no gaps and less than 1 sequence error in 100 kilo base pair (kb)). This finished status is reached by some microbial genomes while plant genomes are in the categories 1 through 4.

The availability of several grass genomes was beneficial for genome-wide studies on genome structure since the genic and intergenic space can be analyzed in all their glory, turning the grasses into an ideal family for comparative genome analysis. This was demonstrated in IBI (2010), where comparison of the genomic sequence of *B. distachyon* with *O. sativa* unveiled a model for chromosome fusion and offers an explanation for the different chromosome numbers in the family of grasses (Paterson *et al.*, 2009).

Table 1.1 A selection of sequenced plant genomes wich are publicly available.
version: the version of the genome sequence at the time of writing; size: genome size; gaps: gaps longer than 5 bp; gene models: number of predicted gene models; n.a.: no information available.

Organism	version	size[Mbp]	gaps	gene models	Source	Reference
A. thaliana	10.0	135	96	33,602	`http://www.arabidopsis.org`	AGI (2000)
A. lyrata	1.0	207	n.a.	32,670	`http://genome.jgi-psf.org/Araly1`	Hu *et al.* (2011)
B. distachyon	1.0	272	1,625	31,029	`http://www.brachypodium.org`	IBI (2010)
O. sativa	7.0	372	203	39,045	`http://rice.plantbiology.msu.edu`	IRGSP (2005)
S. bicolor	1.0	697	6,907	36,338	`http://phytozome.org`	Paterson *et al.* (2009)
V. vinifera	1.0	487	165,717	26,346	`http://www.genoscope.cns.fr`	Jaillon *et al.* (2007)
Z. mays	2.0	2,300	125,338	39,656	`http://www.maizesequence.org`	Schnable *et al.* (2009)
G. max	1.0	975	n.a.	46,430	`http://phytozome.org`	Schmutz *et al.* (2010)
B. rapa	1.1	284	n.a.	66,153	`http://brassicadb.org`	Wang *et al.* (2011)
P. patens	1.6	480	14,910	38,354	`http://cosmoss.org`	Rensing *et al.* (2008)

1–2 Comparative and Evolutionary Genomics in Grasses

Comparing two different genomes usually involves the use of colinearity analysis. Colinearity describes the similar order of genes along chromosomes between two closely related species. In the mid 1980s, restriction fragment length polymorphism (RFLP) markers were developed for applications in plant breeding and genetic research (Gale and Devos, 1998). This resulted in the first genetic maps of cereal crop species. The potential of the RFLP probes to hybridize to highly similar, but not perfectly identical sequences and the low abundance of available markers at that time, stimulated the use of probes from one species for the use in genetic studies in related species.

Thus, the first colinearity across genomes was reported in the late 1980s between tomato and potato (Bonierbale *et al.*, 1988) and between the three diploid genomes of hexaploid wheat (Chao *et al.*, 1988). Soon after, a RFLP-based genetic map was made for the three subgenomes of bread wheat, revealing a high colinearity of marker order between them (Chao *et al.*, 1989). Investigating the genomic relationships of wheat in maize and rice, (Moore *et al.*, 1995b) showed that, despite the divergence of those species approximately 60–70 million year ago (MYA) and their massive differences in genome size, the gene order was still largely conserved along large stretches of the chromosomes.

These early studies also revealed rearrangements between similar genomes, starting the highly productive field of evolutionary genomics. Cross-hybridization of RFLP markers derived from bread wheat with rye (*Secale cereale*) and barley revealed evidence for a few translocations of chromosome arms in the rye genome if compared to the wheat genomes, while most probes showed that the order of the loci was conserved between those three species (Devos *et al.*, 1993; Moore *et al.*, 1995a). The genetic map of rice, the smallest grass genome known at that time,

was divided into linkage groups and aligned against the genetic maps of wheat and maize. The first consensus map for grasses, known today as the 'crop circle', was published in 1995 by Moore *et al.* (1995a), providing the foundation of much of the later research, elaborating and refining the concept and establishing the grasses as a single genetic system. The crop circle indicates that the grasses diverged from a common ancestor and that the gene order seems to be well conserved during evolution even after millions of years, despite chromosomal reorganization and remarkable changes in genome sizes. However, due to the use of only relatively few DNA probes, the genetic resolution of the original crop circle was quite low and did not necessarily reflect the situation at the DNA level. However, it was possible to reconstruct the wheat and maize genome with the rice linkage groups (Moore *et al.*, 1995b). The crop circle was later extended to sugar cane and foxtail millet (Devos, 2005).

The advances in sequence technology and the subsequent decline of costs created a vast amount of sequence information which offered a unique opportunity to investigate colinearity at the molecular level. In fact, already the first studies of genomic colinearity at the sequence level revealed various exceptions, demonstrating that genes were not always found at the expected position and, therefore, the hypothesis of gene movement was formulated (Gallego *et al.*, 1998; Guyot and Keller, 2004). Further comparative analyses of grass genomes revealed many surprising insights into genome evolution. For example, it was found that intergenic regions diverge completely within a few million years (SanMiguel *et al.*, 2002). Only in case of very recent evolutionary divergence, both genes and intergenic regions are still conserved. The finding that the intergenic space is changing at a faster pace than the genic space can be explained by the lower evolutionary pressure for conservation compared to the genes (Petrov, 2001).

1–3 Transposable Elements

Transposable elements (TEs) have been first described as "controlling elements" by Barbara McClintock in 1950 (McClintock, 1950). She was studying maize genetics and observed that some traits lacked a fixed genetic position and had effects on various other genes (McClintock, 1950, 1984). In eukaryotes, the molecular basis of the mechanism involved in these observations was unknown until 1983, when Fedoroff *et al.* (1983) found that the observed effects are due to the DNA-Transposon Activator (*Ac*) and Dissociator (*Ds*) elements. However, the discovery of the repetitive nature of plant genomes was made before. Flavell *et al.* (1974) analyzed the proportion of repeated sequences in 23 plant genomes, amongst them *Triticum aestivum*, *Zea mays* and *Hordeum vulgare*. Since there was no sequence information available yet, this study was done by measuring the reannealing kinetics of denaturated DNA fragments. The results showed that, in average, 80% of each analyzed genome was repetitive.

Later it was discovered that those repetitive sequences are actually mostly TEs. Since repetitive sequences represented a large fraction of a genome without an obvious function, it was proposed that they simply act as spacers between genes. This led to the term "junk DNA" (Ohno, 1972). Due to the inability to transpose and the absence of an evident function, Doolittle and Sapienza (1980) coined the term "selfish DNA".

1–4 Transposable Elements Influence the Genome Structure

Today it is acknowledged that TEs play a central role in genome dynamics. They can influence genome size and affect the expression of genes either directly by inserting into genic regions or indirectly by triggering the response of the host genome to TE activity. TEs have been identified in almost all known organisms (reviewed by Wicker *et al.* 2007a). The term TE is a broad description for various

genomic elements which have one thing in common: they can move in a genome. This ability has an impact on the spatial distribution of genes as well as on the size of the host genome.

The range of genome sizes is very similar in animals and fungi. In mammals, which diverged approximately 70–113 MYA, the genome size averages 3,000 Mbp. In fungi, the described genomes until now range from a few Mbp up to approximately 200 Mbp (Gregory *et al.*, 2007). In reptiles and birds, genomes sizes vary between 1,000–2,000 Mbp (Krishan *et al.*, 2005). In plants, a much broader variation can be observed, even between closely related species. In dicotyledons, genome sizes range from 120 Mbp in *A. thaliana* up to approximately 975 Mbp in *Glycine max* (AGI, 2000; Schmutz *et al.*, 2010). In monocotyledonous plants, which diverged from the dicotyledons approximately 130–240 MYA (Wolfe *et al.*, 1989), an even larger variation in genome size can be found. The smallest monocotyledon genome known until now is that of *B. distachyon* with 273 Mbp, which is already twice the size of *A. thaliana* (IBI, 2010). The genome sizes of rice (389 Mbp, Ouyang *et al.* 2007) and sorghum (697 Mbp, Paterson *et al.* 2009) are considered small compared to the haploid genome sizes of maize and wheat which are 2,500 Mbp and 5,700 Mbp in size, respectively (Schnable *et al.*, 2009; Bennett and Leitch, 1995). Those differences are due to different TE families as they can colonize up to 80% of a grass genome.

Analyses of the organization of genes and TEs in *A. thaliana*, barley, and wheat showed that genes are usually found in clusters, which are separated by long stretches of repetitive sequences, mostly TEs (Barakat *et al.*, 1998; Gill *et al.*, 1996; Künzel *et al.*, 2000; Wicker *et al.*, 2001). In addition, some TEs prefer certain regions as insertion points. For example, it has been shown that centromeric regions, while low in gene content, have a high density of Long Terminal Repeat (LTR) elements (IBI, 2010). In fact, centromeres are almost exclusively colonized by a single group of *Gypsy* elements. These were called centromeric retrotransposons

of maize (CRM) or centromeric retrotransposons of rice (CRR) in maize and rice, respectively (Miller *et al.*, 1998; Nagaki *et al.*, 2005; Wolfgruber *et al.*, 2009; Neumann *et al.*, 2011). In contrast, the small, non-autonomous Miniature Inverted Repeat Transposable Element (MITE) transposons prefer to insert into genes and gene-rich regions (Bureau and Wessler, 1994a).

Interestingly, the number of protein-coding genes, excluding the highly repetitive ribosomal DNA clusters, small nucleolar and small interfering RNAs and conserved non-coding sequences (Freeling and Subramaniam, 2009), is in the fairly narrow range of between 25,000 and 30,000 genes per haploid genome (IRGSP, 2005; Mayer *et al.*, 2011; Paterson *et al.*, 2009). However, the numbers have to be interpreted with caution since gene prediction is still difficult.

1–5 TE Classification

The overwhelming number of TEs and their similarity to genes had made it necessary to categorize and characterize them. Several classification systems arose, as most labs which were working on TEs used their own. Wicker *et al.* (2007a) proposed a unified classification system for all eukaryotic TEs. It is a consensus of the already existing systems and describes guidelines for naming known and newly identified TEs. In this system, two main Classes were defined which are based on the mechanism of transposition. These classes are subdivided into 9 orders and 29 superfamilies. The Class I, which is divided into 5 orders and 17 superfamilies contains TEs which transpose by a "copy-and-paste" mechanism using an RNA intermediate (Figure 1.2). Class II elements use a "cut-and-paste" mechanism without an intermediate (Figure 1.3). They are subdivided into four orders and 12 superfamilies. A three letter code was introduced which reflects the Class, order and superfamily. In addition, the new system can be easily expanded as new types of TEs are identified.

1–6 Class I Elements (Retrotransposons)

Class I elements transpose via a mRNA intermediate which is later reverse transcribed and integrated into the host genome. Therefore, every transposition event creates an additional element (Figure 1.2). Retrotransposons are important forces in genome remodeling and major components of the genome (Kalendar *et al.*, 2000; Schulman *et al.*, 2004; Vicient *et al.*, 1999). In plant genomes, the LTR elements are the most abundant Class I members. They resemble retroviruses in their structure and life cycle but are not known to be infectious (Kumar and Bennetzen, 1999). Like retroviruses, these genomic components replicate by a cycle of transcription of the integrated copies, as if they were cellular genes, followed by translation of their encoded products and reverse transcription of the RNA into cDNA. The two LTRs flank an internal domain which, in fully functional elements, encodes the proteins necessary for their own replication and are needed for retrotransposition (Schulman and Kalendar, 2005). These proteins are present as two main Open Reading Frames (ORFs). These are: Gag, specifying the structural protein forming the nucleocapsid and Pol, encoding the enzymatic functions. Pol is a polyprotein and is auto-processed by its aspartatic proteinase (AP) domain. It contains reverse transcriptase (RT) and RNAseH, a bifunctional polypeptide carrying out reverse transcription and integrase (INT), which inserts the new LTR retrotransposon copy into the genome (Suoniemi *et al.*, 1998). In some elements such as *BARE-1*, the two ORFs are fused into one (Manninen and Schulman, 1993).

Gypsy and *Copia* are the main superfamilies. They differ in the order and sequence domains of the encoded ORFs. Despite their abundance, the copy numbers of individual families varies strong. Most families are found in low or modest copy numbers (1 – hundreds). Still, some families successfully colonized their host genome, representing a large fraction of the genome. In the Triticeae genomes,

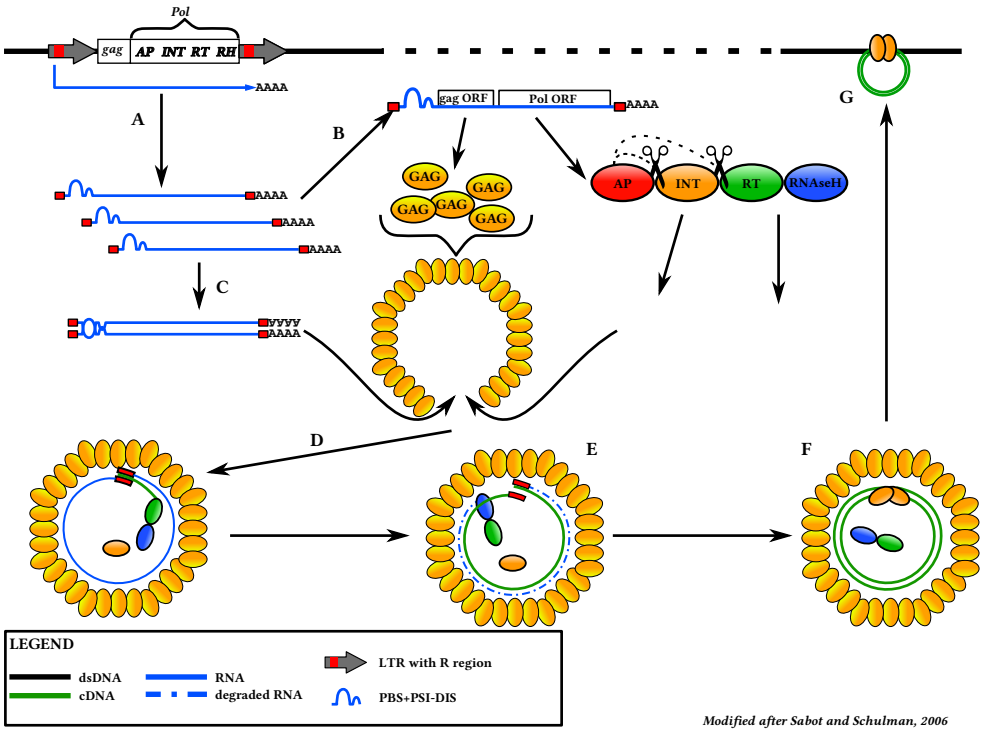


Figure 1.2 The life cycle of a LTR retrotransposon explained with a *Copia* element. Elements are colored as described in the legend. (**A**) Transcription of the mRNA. (**B**) Synthesis of the *gag* and *Pol* acpORF. *Pol* is further cleaved into the the dual protein reverse transcriptase (RT) / RNAseH and integrase INT by the aspartatic protease (AP). GAG proteins are assembled into the virus like package (VLP). (**C**) RNA packaging through dimerization, using a "kissing-loop" mechanism directed by DIS recognition. (**D**) Synthesis of the first cDNA strand inside the VLP. (**E**) Degradation of RNA and synthesis of second cDNA strand. (**F**) INT binds to the end of the double strand cDNA. (**G**) Insertion of the newly copied element into a new location by introducing a double strand break.

BARE1, *WIS* and *Angela* elements account for more than 10% (Vicient *et al.*, 1999; Kalendar *et al.*, 2000; Wicker *et al.*, 2009b). In barley, 50% of the genome is made up of only 14 TE families, of which 12 are LTR retrotransposons (Wicker *et al.*, 2009b).

The factors responsible for successful colonization have not been yet identified. Stress condition may trigger genome expansions, as it was shown that some abundant families get activated upon drought or UV light (Kalendar *et al.*, 2000; Ramallo *et al.*, 2008). On the other hand, other stress-induced retrotransposons are still only found in low copy numbers (Grandbastien *et al.*, 2005). This indicates that either some families have a predisposition to successfully transpose in high number or/and that evolutionary forces selectively keep the copy numbers of some families low.

The characteristics of LTR retrotransposons make them a favorite tool to analyze genomes. The LTRs from each element are identical at the time of insertion. After insertion, both LTRs accumulate mutations independently according to the molecular clock and by comparing the two LTRs of a given element, the insertion time can be estimated (SanMiguel *et al.*, 1998). Additionally, transposition leaves a copy of the element at the original location, allowing linkage between insertion time and the genome position of the original element. Previous studies analyzed the insertion times and reported that particular families of LTR retrotransposons are active in particular epochs which may span several hundreds of thousands of years (Pereira, 2004; Wicker and Keller, 2007). What triggers these "waves" of activity of long periodicity is not known yet.

1–7 Class II Elements (DNA-Transposons)

Class II elements transpose via a DNA intermediate, hence the name DNA-Transposon. In contrast to Class I elements, where a copy at an element is transposed, DNA-Transposon elements transpose the excised element (Figure 1.3). The known 12

superfamilies of DNA-Transposon elements are found in virtually all eukaryotes (Wicker *et al.*, 2007a; Ueki and Nishii, 2008; Oosumi *et al.*, 1996; Feschotte and Mouchès, 2000; Smit and Riggs, 1996). The most common characteristics of a DNA-Transposon is the presence of Terminal Inverted Repeats (TIRs), a DDE transposase and a length of a few kb (Fedoroff *et al.*, 1983; Wicker *et al.*, 2003a; IBI, 2010). The exceptions are elements from the superfamilies *CACTA*, *Harbinger*, *Crypton*, *Helitron* and *Maverick*. The former two encode an additional ORF (ORF2), while the latter three have a completely different structure with several ORFs (Wicker *et al.*, 2007a). In some families of the *CACTA* superfamily the ORF2 has been described to support the excision and transposition (Frey *et al.*, 1990). However, its function is not yet completely clear.

The Subclass 2 contains more exotic elements. *Crypton* contains only a tyrosin recombinase and no TIRs. *Helitrons* are also missing the TIRs and its tyrosin recombinase builds a bi-protein with a helicase. In addition, in plants a replication protein A is found in complete *Helitrons*. The *Maverick* elements have TIRs and harbor four different proteins, a C-Integrase, a packaging ATPase, a cystein protease and a DNA polymerase B.

The transposition mechanics of DNA-Transposons are not yet completely understood. The excision of most DNA-Transposons is catalyzed by a DDE type transposases. The name derives from the catalytic triad of negatively charged amino acids `DD[E|D]` which bind divalent metal ions and are located in an RNAseH-like fold. The first structure of an eukaryotic transposase from the DNA-Transposon *Mos1*, a *Mariner* element from *Drosophila melanogaster*, was recently published (Richardson *et al.*, 2009). The transposase directs the three crucial steps in the transposition and does not require an external energy source (reviewed in Mizuuchi 1997, Rice and Baker 2001). It has to recognize and pair the the specific TIR sequence of a DNA-Transposon, forming a synaptic complex. Then, the DNA at both ends

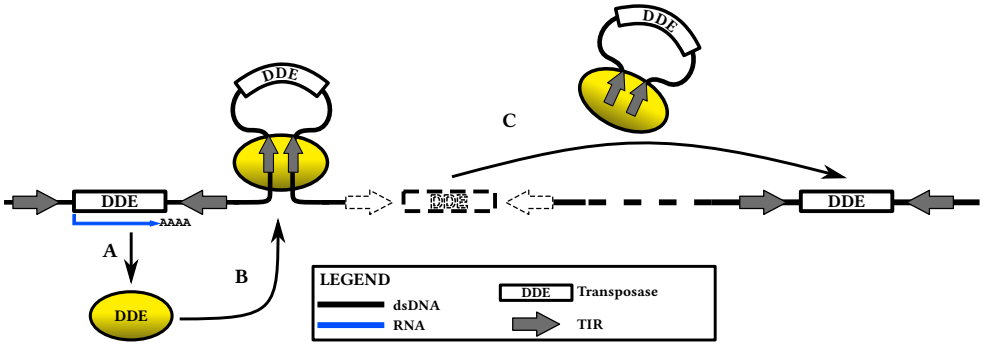


Figure 1.3 The life cycle of a DNA transposon.
The different parts are described in the legend. (**A**) Transcription and translation of the transposase mRNA. (**B**) Binding of the transposase (yellow) to the TIRs of the TE, creating a hairpin-like structure. (**C**) Excision of the whole TE and insertion into the new location. The excision as well as the insertion lead to a DSB. The previous location is indicated by the dashed outline of the element.

of the element has to be cut, releasing the 3′-OH ends of the DNA-Transposon. This step uses water as attacking nucleophile. The final step is the insertion into the new location, this time using the 3′-OH groups as attacking nucleophile. Crystallographic analysis of the transposase *Mos1* indicate that the transposase acts as a dimer and recognizes the TIRs (Richardson *et al.*, 2009). The resulting synaptic complex, called paired-end complex (PEC) and resembling a hairpin like structure, is excised and transposed to the new location. Additionally, some previous studies have shown that the excision of some *Mariner* elements does not produce blunt end double strand breaks but creates small 3′ overhangs by leaving behind a few nucleotides of the terminal inverted repeat of the element (Dawson and Finnegan, 2003; Yang *et al.*, 2006; Robert and Bessereau, 2007; Richardson *et al.*, 2009). The mechanics for insertion are still not known.

In contrast to Class I elements, Class II elements account only for 8% to 13% of a plant genome (Wicker *et al.*, 2009b; IBI, 2010; Schnable *et al.*, 2009; Paterson *et al.*, 2009). However, DNA-Transposons usually have high copy numbers, e.g. in *B. distachyon* 29,630 DNA-Transposons were identified, 143,000 in *Z. mays* and

294,937 in *G. max* (IBI, 2010; Schnable *et al.*, 2009; Schmutz *et al.*, 2010). To explain such large numbers with a mechanism which does not leave a copy at the original location, a model for the replication of DNA-Transposons has been proposed. In this model, the DNA-Transposon transposes during DNA replication. When the replication fork has passed a DNA-Transposon, an element on the newly synthesized strand transposes in front of the proceeding fork. This will lead to an additional copy on the lagging strand. However, this model could not be yet experimentally verified. While retrotransposons are mostly found in gene poor regions, DNA-Transposon are usually found in gene-rich regions and introns (Bureau and Wessler, 1994a).

1–8 The Formation of the Target Site Duplication

Insertion of TEs goes usually hand in hand with the creation of a target site duplication (TSD) at the insertion site and occurs in both classes of TEs. Except for the TEs of the order DIR (Class I), Crypton and Helitron (both Class II), insertions of a TE creates a TSD. TE insertion causes a DSB which is initiated through a staggered cut of the DNA at the insertion site by the transposase, similar to a restriction enzyme. Upon insertion, the complementary sequence of the overhangs which are flanking the freshly inserted element, get polymerized and thereby duplicated. This creates a copy of the target site on both sides of the element (Figure 1.4). The size of the TSD can range from 2 bps up to 16 or more bp and is specific for each superfamily, e.g the TSD for *Mariner* elements is the dinucleotide TA while the TSD for *Mutator* elements is between 9–12 bp long and starts with GC (reviewed by Wicker *et al.* 2007a). Because the TSDs for an inserted element are identical and border it, they are used as diagnostic motif in TE analysis to assure the completeness of an identified element.

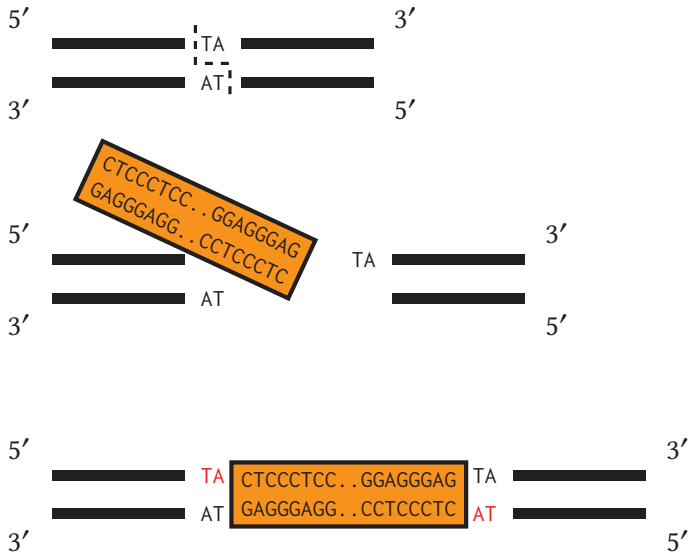


Figure 1.4 Creation of a TSDs explained with a *Mariner* element.
The *Mariner* element is drawn as orange box with its TIRs written inside. It is not drawn to scale and the transposase is not indicated for clarification. Black lines represent DNA with the orientation indicated by 5' and 3', respectively. The dashed line depicts the staggered cut of the transposase prior to insertion. Newly synthesized TSDs are indicated in red.

1–9 Autonomy and Parasitism within the Genome

TEs are present in two variations: autonomous and non-autonomous. Autonomous TEs harbor all proteins needed for transposition while non-autonomous elements represent partial deletion derivatives which have lost their ability to transpose (Figure 1.5). For non-autonomous DNA-Transposon elements, historical reasons lead to the distinction of two classes: (i) non-autonomous elements and (ii) Miniature Inverted Repeat Transposable Element (MITE). The difference is very fuzzy, since by definition the former elements have a truncated version of the transposases while the latter have lost the transposase completely (Figure 1.5, reviewed by Casacuberta and Santiago 2003). MITEs were first described in grasses by Bureau and Wessler (1994b, 1992) but have been identified in numerous other species, e.g. fungi (Yeadon and Catcheside, 1995), mosquitoes (Tu, 1997), beetles (Braquart *et al.*,

1999), Xenopus (Unsal and Morgan, 1995), humans (Smit and Riggs, 1996) and fish (Izsvák *et al.*, 1999). Yang *et al.* (2006) demonstrated in yeast the transposition of the non-autonomous *Osmar5* element. They expressed the complete transposase on a separate vector and observed that it could mobilize the non-autonomous element which was located on a different vector. Therefore, we use the term non-autonomous in conjunction with DNA-Transposons to describe all elements which have no or only a partial transposase gene but retained their TIRs.

Similarly, non-autonomous Class I elements have been described (Sabot and Schulman, 2006). Since retrotransposons contain several ORFs, non-autonomous elements arise from premature stop codons, frame shifts or partial deletions of the coding sequence (CDS). Those are generated due to high error rates in the steps of transcription and reverse transcription (Preston, 1996; Boutabout *et al.*, 2001). The resulting non-autonomous elements are classified in three groups: *LARD* (large retrotransposon derivative, Kalendar *et al.* 2004), *TRIM* (terminal repeat in miniature, Witte *et al.* 2001) and *Morgane* (Sabot *et al.*, 2006). The *LARD*s consist of long LTRs on both sides which flank a long and conserved internal domain without protein coding capacity. *TRIM*s have short LTR and the internal domains contain only signals for the reverse transcription. *Morgane* elements harbor small and non-functional remains of the *Pol* ORF (Figure 1.5). In contrast to some non-autonomous DNA-Transposons which were activated in *trans* by an autonomous element, no *trans*-activation has been demonstrated for non-autonomous retrotransposon elements (Yang *et al.*, 2006).

1–10 TEs Influence the Epigenome of the Host

The epigenome describes the sum of modifications which indirectly influence the coding capacity of the genome, either by modulating post-transcription levels, histone modification and methylation (Kouzarides, 2007). The expression and activity

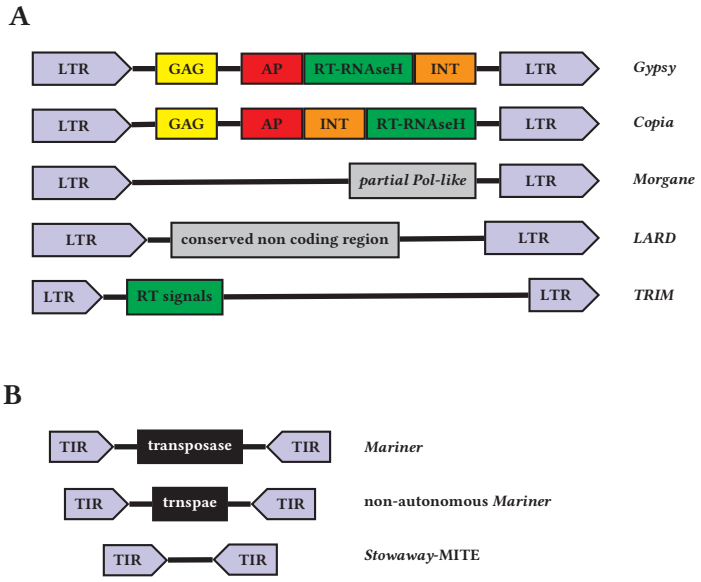


Figure 1.5 Schematic comparison of autonomous and non-autonomous TEs. (A) Typcial non-autonomous Class I elements compared to the autonomous LTR retrotransposons *Copia* and *Gypsy*. LTR: Long Terminal Repeat. GAG: GAG protein. AP: aspartatic proteinase. RT-RNAseH: the bi-protein reverse transcriptase and RNAseH. INT: integrase. AP, RT-RNAseH and INT form the *Pol* ORF. RT signals indicates the signals for reverse transcription found in *TRIM*s. (B) Non-autonomous Class II elements explained with the *Mariner* superfamily. TIR: Terminal Inverted Repeat (TIR). The black box represents the the transposase where internal deletions are indicated by missing letters.

of TEs can be controlled by the plant. Similar to genes, which can be inactivated by methylation, TEs can also be inactivated by epigenetic pathways (Lisch, 2009). In this process, the chromatin structure is modified to enclose the targeted sequences in different states of nucleosome packing. This allows to modulate the activity of the RNA polymerase on a so called "active" sequence. It has been shown that the control of the chromatin state is linked to small interfering RNA (siRNA) silencing, nucleosome remodeling DNA methylation and histone modification (reviewed by Holmquist and Ashley 2006). Since TE code for their own proteins, those transcripts can get recognized and are used in controlling the expression of TEs, resulting in a lower or even stalled TE activity.

In plants, this reduction in proliferation is controlled by epigenetic silencing, a mechanism where TE are methylated due to a combination of siRNA and proteins (Lisch, 2009; Matzke *et al.*, 2009). This post-transcriptional gene silencing produces small and specific RNAs that target the DNA for methylation, which in turn interacts with mechanisms involved in chromatin methylation. The mutation of DDM1, a chromatin remodeling factor in *A. thaliana*, led to a genome wide demethylation which in turn led to a burst in transposition of retrotransposons and a DNA-Transposons (Tsukahara *et al.*, 2009). However, this regulation influences the expression of genes which are nearby TEs, e.g. MITEs and *TRIM*s (Kalendar *et al.*, 2008; Huang *et al.*, 2008; Menzel *et al.*, 2012). In rice, Lu *et al.* (2012) observed that the genes associated with MITEs have a significantly lower expression. Similar, *Cassandra TRIM*s or *Tos-17* in rice are often associated with genes, indicating that a high proportion of genes is exposed to the influence of TEs. Since TEs can be activated due to stress conditions in plants and influencing the epigenome by triggering the siRNA pathways with their transcripts, there is a potential link between stress and the dynamics of the epigenome.

1–11 TEs as Driving Force for Genome Evolution

TE activity has an influence on gene activity and genome evolution. Therefore, it is important to understand which TEs are dynamic, their mode or replication as well as the interaction between different elements and the genome. Studying these mechanisms will improve our knowledge how genomes change over time and which evolutionary mechanism are driving those changes. The continuing improvement in sequencing technology is releasing a flood of genome sequence and expression data in a unprecedented resolution. The analysis and mining of this huge amount of data requires bioinformatic approaches to reveal new hypothesis which must then be tested experimentally.

TEs contain only the amount of information which is needed for their own transposition and recruit the translation and replication machinery from their host. This aspect lead Doolittle and Sapienza (1980) to the hypothesis that the only function of TEs is their survival in the genome, but at the same time, this "raw material" can have adaptive values later in evolution.

1–12 TE Contribution to Gene Evolution

All jawed vertebrates depend on their adaptive immune system for recognition and defense of numerous pathogens. The antigen receptors, which are highly specialized for a specific ligand, are not encoded genetically since the diversity would be limited. Therefore, a defense system evolved which uses a combinatorial solution (Tonegawa, 1983). The immune cells contain three variable segments, the V (variable), D (diversity) and J (joining segments). The specificity of the immune cell is created by combining those three segments into one exon which encodes the unique recognition site for a specific antigen. The assembly of such a recognition site is called V(D)J recombination. The recombination is facilitated by two proteins, RAG1 and RAG2 (Oettinger *et al.*, 1990). Since the V(D)J recombination show

similarities to the transposition of DNA-Transposons, it was believed that this mechanism evolved from a ancient transposase. Indeed, Kapitonov and Jurka (2005) found that the core regions of RAG1 is shared by the DDE transposase from the DNA-Transposon *Transib*. Therefore, they suggested that the RAG1 derived form an ancient *Transib* element.

Another example where TEs have been domesticated by the host genome is the maintaining of the telomeres in *D. melanogaster*. The linear DNA in eukaryotes cannot be completely duplicated by the DNA polymerase due to the linear form and the semi-conservative replication mechanism (Watson, 1972). The telomeric ends consist of a characteristic and simple tandem repeats and are similar in several species, e.g. `TTAGGG` in vertebrates, `TTAGG` in insects and `TTTAGGG` in plants (Pardue and Debaryshe, 2011). The conservation and maintenance of telomeric ends, also called telomere length homeostasis, is crucial. Short telomeres trigger DNA damage checkpoints which lead to cell senescence. In addition, those telomeric caps add an additional safety to differentiate the telomeric ends from chromosome breaks (Levis, 1989). In eukaryotes and some other species, telomeric length maintenance is based on the telomerase activity, a specialized reverse transcriptase with a complementary telomeric repeat sequence (Greider and Blackburn, 1985, 1987). This sequence is used as template for the reverse transcription (Yu *et al.*, 1990). However, *D. melanogaster* lacks a telomerase. Instead, the telomeres are ordered arrays of three retrotransposons (*HeT-A, TART, TAHRE*). Their transcribed copies serve as templates for reverse transcription which maintains the length of the telomeres (Pardue and DeBaryshe, 2011).

1–13 Genomic Turnover

TEs are not only inserted, but also removed from the genome. For retrotransposons two ways are known: one mechanism involves the recombination of the LTRs,

which removes the whole element except one recombined copy of an LTR, which is then called a "solo LTR". This mechanism is based on normal homologous recombination and can occur between LTRs which show a high degree of similarity and possibly between two direct repeats which are found a few kb apart on the same DNA strand (Vicient *et al.*, 1999; Ma *et al.*, 2004; Devos *et al.*, 2002). The interacting repeats, e.g. LTRs from the same element of two adjacent elements, form a loop-like structure. The sequence between the pairing repeats is lost. For DNA-Transposons no such mechanisms have been described yet. Another way of DNA removal is through small deletions and truncations by a mechanism called illegitimate recombination (IR). Despite the fact that the molecular mechanism for IR is so far unknown, it is believed that this drives the the fast turnover of intergenic sequences (Devos *et al.*, 2002; Wicker *et al.*, 2003a).

Demographic changes in the TE population correspond to waves of gain and loss. In some cases, under the assumption of a constant removal rate of LTR elements, a half life for each family can be calculated. For example, the half-life for *Copia* elements in rice is about 790,000 years (Wicker and Keller, 2007). Studies in closely related species have shown that the intergenic space between the genes is very dynamic. Genes tend to be found in colinear positions, while the intergenic space has been almost totally replaced. SanMiguel *et al.* (2002) predicted that after 10–14 million year (Myr) the intergenic space at a locus will have changed totally. However, using an updated substitution rate of 3.1×10^{-8} substitutions per site per year (as proposed in Ma and Bennetzen 2004), those numbers need to be adjusted to 5–7 Myr. Likewise, newer studies showed that the time of intergenic turnover is actually much faster, e.g., in rice, most LTR insertion happened in less than 6 Myr after divergence (Ma *et al.*, 2004). Comparisons between *B. distachyon* and *B. sylvaticum* showed that the intergenic space has already underwent extensive changes after only 1.7–2.4 Myr (Buchmann *et al.* 2012, part of this study, see Chapter 3).

While all TEs cause a DSB upon insertion, DNA-Transposons also cause a DSB upon excision which has to be repaired by the host cell. In addition, the close vicinity of DNA-Transposons to genes can lead to an indirect influence on gene movement and duplication. The repair of such a DSB using the sequence dependent strand annealing (SDSA) pathway can duplicate a gene by using it as a repair sequence (Wicker *et al.*, 2010). Deletion occurs if the DSB is repaired by the simple sequence annealing (SSA) pathway. This shows that the mobility of DNA-Transposons has a considerable and, until now, neglected impact on the intergenic turnover.

1–14 Aim of the Thesis

The aim of this thesis is to analyze the the role of TEs in plant genome evolution and the interaction of the host genome with its mobile DNA fraction. The availability of the recently sequenced genome of *B. distachyon* and a BAC-library of its close relative *B. sylvaticum* allowed us to investigate the evolutionary processes which occur shortly after divergence from the common ancestor. The high quality of the genomic sequence from *B. distachyon* allows to investigate biological mechanisms which occur between genes in the intergenic space, mostly TEs.

A first project was to characterize the *CACTA* population in the newly sequenced *B. distachyon* by identification and classification of individual families. The genome size of 273 Mbp from *B. distachyon* required the development of a pipeline to automate as many steps as possible while maintaining a high quality of analysis.

In a second project, we used five orthologous loci from *B. sylvaticum* and *B. distachyon*, spanning 1 Mbp in total, to analyze the impact of DNA-Transposons on the sequence composition after species divergence. This required the estimation of the divergence time and detailed annotations of genes and TEs. New programs and approaches for sequence analysis with a strong focus on TEs were written and tested during the projects.

Analysis of *CACTA* Transposons in Grasses*

2

Abstract

CACTA transposases show different exon numbers among *CACTA* families. Despite this difference, the transposase is always a key enzyme for transposition and replication. To derive a model which could explain the observed differences in exon numbers, we compared 44 different transposases from six different species. To build our dataset, we performed an *in silico* approach to annotate and characterize *CACTA* families in the recently sequenced *Brachypodium distachyon* genome. We identified 14 *CACTA* families with 1,998 elements in total which covered approximately 3% of the genome. Three families were classified as non-autonomous due to the lack of a transposase. We used the 10 transposases from the remaining putative autonomous elements and expanded the dataset with 34 transposases from *Arabidopsis thaliana*, *Petunia hybrida*, *Zea mays*, *Triticum aestivum* and *Oryza sativa* which were identified in *PTREP*. The phylogenetic analysis of the protein sequences from the transposases indicated that the main lineages diverged already before the monocotyledon and dicotyledon divergence. Based on the analysis of conserved exon/intron boundaries between the transposases, we propose that the ancient transposase in grasses contained at least four exons. We developed a model for the formation of different exon numbers through differential loss of introns. In one

*Part of the presented data has been published in IBI (2010)

case we identified an event which removed all introns for a specific group of *CACTA* transposases.

2–1 Introduction

The *CACTA* superfamily belongs to the Class II of transposable elements, proliferating with a "cut and paste" mechanism. In contrast to Class I elements, which transpose via an RNA intermediate and therefore copy the original element, Class II elements transpose the original element. *CACTA* elements were named after their characteristic CACTA motif at the end of their Terminal Inverted Repeats (TIRs). The first *CACTA* element described at the molecular level was *En-1* in *Zea mays* (Pereira *et al.*, 1986) which is the autonomous element of the Suppressor-mutator (*Spm*) family. Active *CACTA* elements have been described in *Arabidopsis thaliana* (*CAC1*, Miura *et al.* 2001), *Petunia hybrida* (*Psl*, Snowden and Napoli 1998), *Daucus carota* (*Tdc1*, Ozeki *et al.* 1997), *Sorghum bicolor* (*Candystripe1*, Chopra *et al.* 1999), *Antirrhinum majus* (*Tam-1*, Nacken *et al.* 1991) and *Ipomoea nil* (*Tpn-1*, Inagaki *et al.* 1994). In addition, Ueki and Nishii (2008) reported the finding of *Idaten*, a *CACTA*-like element in *Volvox carteri*.

A full-length *CACTA* element consists of both TIRs and two Open Reading Frames (ORFs), of which one encodes a transposase and the second a protein of unknown function, called ORF2. In addition, several subterminal repeats are found between the ORFs and the TIRs (Figure 2.1). The transposase is the key transposition enzyme and binds to the TIRs during excision, creating a 3 base pair (bp) target site duplication (TSD) (Lewin, 1997). The catalytic center of the transposase is the acidic triad known as the "DDD/E" motif (Yuan and Wessler, 2011). The function of the ORF2 protein has been determined in specific *CACTA* families to bind to the sub-terminal repeats, supporting the excision and transposition (Frey *et al.*, 1990). The length of the TIRs can reach 30 bp, of which the first and last 5 bp consist

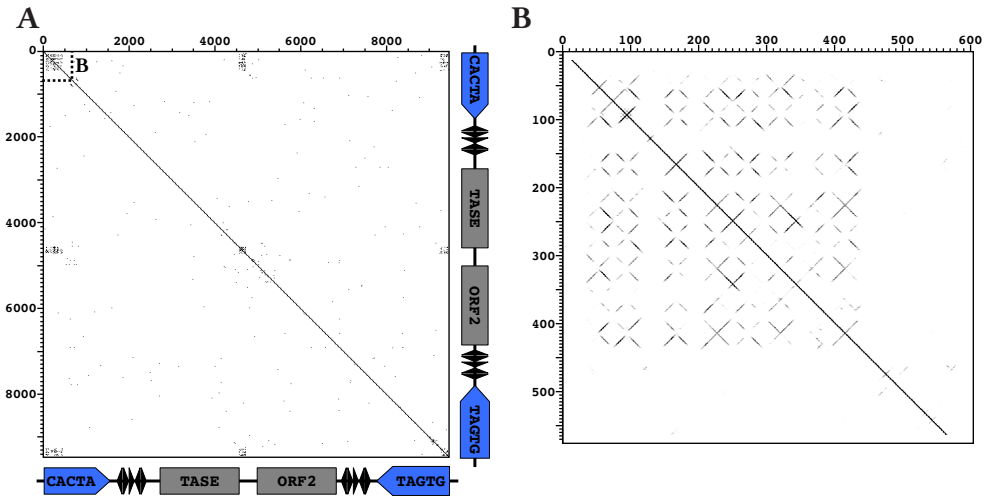


Figure 2.1 Characteristics of a *CACTA* DNA-Transposon.
(A) Dot-plot of CACTA_A from *B. distachyon* versus itself to show the characteristics of an *CACTA* element. The cartoons at the bottom and right depict the characteristics of a typical *CACTA* element (not to scale). Blue arrows: TIRs; Grey boxes: ORFs; Black triangles: sub terminal repeats. The region marked B is shown at a higher zoom in (B).
(B) Close up of region B in (A), showing the characteristic sub terminal repeat pattern.

of the `CACTA` and `TAGTG` motif, respectively. Unfortunately, those motifs are the only sequence which is usually conserved between different families, therefore rendering the identification of new elements based solely on the TIR unfeasible. The subterminal repeats, usually between 10–20 bp long and found in direct and indirect orientations, are also not conserved between families. Therefore, the motif and the presence of the transposase and ORF2 proteins are the only specific features for the identification of new *CACTA* families.

Despite the fact that Class II elements usually do not account for the large genome sizes found in grasses, *CACTA* families show a high abundance, like *Tpo1* in *Lollium perenne* (ryegrass) and *Caspar* in the Triticeae, which contributed considerably to the expansion in genome size (Langdon *et al.*, 2003; Wicker *et al.*, 2003a, 2009b). In addition, *CACTA*s influence the evolution of the host genome (Bennetzen, 2000). In *Glycine max* (Soybean), *CACTA* elements can influence the flower color and capture host genes (Xu *et al.*, 2010; Zabala and Vodkin, 2007). Likewise, host gene capture has been described in *Brassica oleracea* and *Sorghum bicolor* (Alix *et al.*, 2008; Paterson *et al.*, 2009). *CACTA* elements can be associated with regulatory elements of genes, therefore possibly influencing gene expression (Wicker *et al.*, 2003a, 2005).

The results presented in this chapter focus on *CACTA* elements and include the work for BRAC (*Brachypodium* Repeat Annotation Consortium, IBI 2010), where we were responsible for the characterization and annotation of *CACTA* elements. We identified and characterized 14 families with a total of 1,998 elements and covering approximately 3% (8 mega base pair (Mbp)) of the *B. distachyon* genome. As a follow-up, we performed a comparative analysis between 44 *CACTA* transposases from the six species *B. distachyon*, *Z. mays*, *A. thaliana*, *Oryza sativa*, *Triticum aestivum* and *Petunia hybrida*. Our study focused on the number of exons of the *CACTA* transposase genes. We analyzed the structural differences between different *CACTA* families across the six species, thereby gaining an insight into the evolutionary dynamics of *CACTA* transposase genes.

2–2 Results

To identify *CACTA* families in *Brachypodium distachyon*, we performed an *in silico* search of the genomic sequence. We set up a pipeline with four steps, of which the first, second and last were automated while the third one consisted of a semi-automated step. The experimental procedures are described in detail here in the results section because there is no widely used protocol for *CACTA* identification and many methods had to be developed. We added a summary of the used methods and programs at the end of the chapter.

Extracting Putative *CACTA* Elements from the *B. distachyon* Genome

The identification of putative elements was done with the `Perl` program `cacta.pl`. This program searches for the characteristic Terminal Inverted Repeats (TIRs) of a *CACTA* element with the highly conserved motif `CACTA` on the 5′ end and `TAGTG` on the 3′ end of the element. The length of the elements can differ, but we were screening for autonomous elements which are several kilo base pairs (kb) in length. Therefore, after finding the 5′ motif, the program searched for the corresponding 3′ motif in a distance between 8–12 kb. The minimal length of 8 kb ensured that mostly autonomous elements were recognized. In maize and sorghum the average length of *CACTA* elements was approximately 5 kb, including truncated elements. We wanted to reduce the number of truncated elements and therefore screened for longer elements (Paterson *et al.*, 2009; Schnable *et al.*, 2009). Non-autonomous elements are usually smaller and were found later using already characterized *CACTA* sequences as query in `BLASTN` alignments against the *B. distachyon* genome. Upon finding a corresponding 3′ motif, the program compared the three nucleotides before the 5′ and after the 3′ motif to verify the target site duplication (TSD). If no 3′ motif in 12 kb or no matching TSD was identified, the search was aborted. To avoid storing of the whole genome sequence in the memory, sliding windows of

20 kb were analyzed which were shifted by 5 kb.

Screening of Putative *CACTA* Elements for Transposase and ORF2

The first step lead to the identification of 5,073 putative *CACTA* elements, of which numerous were false positives due to the fact that the characteristics motifs can also be present by chance. The putative elements were screened for the presence of a transposase and ORF2 by BLASTX searches against the proteins in PTREP, reducing the data set to 173 putative elements.

Manual Check of Putative *CACTA* Elements

The 173 putative elements were manually check by dot–plot alignments for the presence of the ORF2 and intact ends which are characterized by several direct and indirect repeats. Once a complete element was identified, copies with ≥80% similarity on the DNA level were removed from the set of putative elements and extracted from the genomic sequence to obtain a consensus sequence.

Defining a *CACTA* Family

The consensus sequence was considered a representative of a new family if it had not been identified in a previous run and showed <80% DNA identity. The new consensus sequences were used in BLASTN searches against the genomic sequence to mask and retrieve the individual elements from each family. We considered all *CACTA* elements which showed ≥80% similarity to the consensus sequence as family member. We identified 14 *CACTA* families in total and named them *CACTA_A* through *CACTA_N* (Figure 2.2A, Table 2.1).

For the three families *L, M* and *N* no transposase and ORF2 were identified, an indication that these families were non-autonomous. In addition, we identified a family with TIRs that showed high similarity with the *CACTA_A* consensus sequence but no similarity with other families. This family was also missing the transposase and ORF2, very likely representing a non-autonomous subset of the *A*

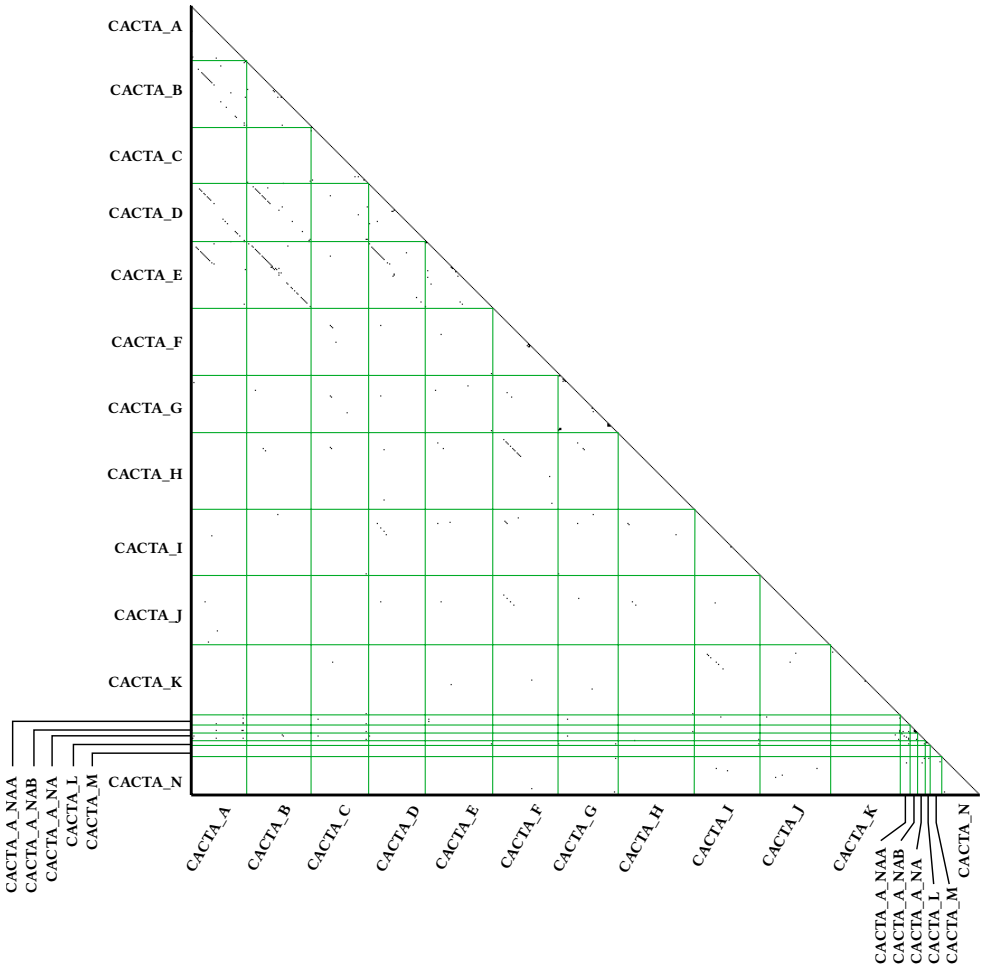


Figure 2.2 Dot-plot of the DNA consensus sequences from all identified *CACTA* families in *B. distachyon* to show the high variability as only few CDS regions show conserved parts.
The 14 DNA consensus sequences plus three subfamilies from *CACTA_A* were aligned against itself. Green lines separate the individual consensus sequences which are labeled *CACTA_A* to *CACTA_N*. The non-autonomous families are *CACTA_L*, *CACTA_M*, *CACTA_N* and the non-autonomous subfamilies from *CACTA_A* are *CACTA_A_NA*, *CACTA_A_NAA* and *CACTA_A_NAB*.

Table 2.1 Summary of the identified 14 *CACTA* families in *B. distachyon*.
#: number of elements in this family; tase: number of predicted exons for the transposase; ORF2: number of predicted exons for the ORF2; cons: length of the consensus sequence; coverage: total coverage of all elements in this family; ∅ length: average length for an element in this family; Σ: total; n.a.: unknown exon number; none: not found

	family	tase	ORF2	cons [bp]	#	coverage[bp]	∅ length[bp]
Auto	*A*	1	4	9,419	479	2,905,452	6,065
	B	1	4	11,807	379	1,910,447	5,040
	C	2	n.a.	9,711	28	107,618	3,843
	D	1	4	10,430	132	564,101	4,273
	E	1	n.a.	11,787	229	977,223	4,267
	F	2	n.a.	11,681	5	38,819	7,763
	G	2	5	10,049	261	728,207	2,790
	H	2	n.a.	13,655	18	93,236	5,179
	I	3	8	11,656	25	87,000	3,480
	J	3	n.a.	12,226	51	146,166	2,866
	K	3	n.a.	12,436	16	67,764	4,235
Σ					1,623	7,626,033	4,698
Non–Auto	*A_NA*	none	none	1,309	260	404,585	1,556
	L	none	none	796	77	63,359	822
	M	none	none	2,654	28	46,646	1,665
	N	none	none	6,115	10	29,275	2,927
Σ					375	543,865	1,450
Σ					1,998	8,169,898	4,089

family and not an own family per se. We named this family *A_NA*. In the *A_NA* group two additional subfamilies were identified and labeled *CACTA_NAA* and *CACTA_NAB* (Figure 2.2B). The 11 putative autonomous families (i.e. families for which we annotated complete transposase genes) have 1,623 copies with an average length of 4,698 base pair (bp) and cover 7.6 mega base pairs (Mbp) of the genomic sequence (93% of all *CACTA* sequences). The non-autonomous families have 375 copies with an average length of 1,450 bp and cover 543.865 kb (7% of all *CACTA* sequences, Table 2.1). In total, we identified 1,998 *CACTA* elements which cover 8.1 Mbp, approximately 2.9% of the *B. distachyon* genome (Table 2.1).

Different Exon Numbers in the Transposase and ORF2 Genes in *B. distachyon CACTA* Families

We further analyzed the structural organization of the *CACTA* elements by annotating the transposase and the ORF2 on the consensus sequence of the 11 putative autonomous families. The transposase exon/intron boundaries were not always precisely known since transcriptome data is scarce. Therefore, the annotations were mostly based on aligning proteins of known transposases against the consensus DNA sequence. The transposase was annotated relatively easily in all consensus sequences except for *CACTA_C*, were no clear exon/intron boundaries were identified. Therefore, we removed *CACTA_C* from further analyses. The high variability of the ORF2 made its annotation difficult. Nevertheless, we annotated the ORF2 for the five families *A, B, D, G* and *I*. The remaining families showed no similarities to known ORF2 sequences, neither on the DNA nor the protein level and gene predictions did not return significant results.

The comparison of the exon number of the annotated transposases and ORF2 identified three distinct groups in *B. distachyon*. The families *A, B* and *D* have a transposase with one exon and an ORF2 with four exons while family *G* has

a transposase with two exons and an ORF2 with five exons. Family *I* has an transposase with three exons and an ORF2 with eight exons (Figure 2.3). Since the transposase was annotated in all families and is crucial for transposition, we decided to focus on it for further analysis.

To investigate the structural diversity among the *CACTA* elements, we included 35 additional elements from *Arabidopsis thaliana*, *Zea mays*, *Triticum aestivum*, *Oryza sativa* and one from petunia (*Petunia hybrida*), creating a final dataset of 44 *CACTA* transposases (Table 2.2). The petunia element, which is very divergent (and probably represents an ancient *CACTA* lineage) was chosen as outgroup.

We annotated the exons for each transposase using dot–plot alignments of the coding sequence (CDS) against the transposase proteins which were identified by `BLASTX` searches versus the PTREP database (Figure 2.3). The number of exons in the investigated *CACTA* elements ranged from 1 to 5. The majority are two groups of 14 elements each with 1 or 4 exons. In 9 elements we annotated 3 exons while only 3 elements had 5 exons (Table 2.2).

Figure 2.3 ***(following page)*** The three different *CACTA* configurations identified in the 14 families of *B. distachyon*, based on the exon number in the transposases and ORF2. The x-axis indicates the DNA consensus sequence. The sequences on the y-axis are CDS sequences of the transposase (left) and ORF2 (right), indicated by CDS_{tase} and CDS_{ORF2}, respectively. The arrow indicates the orientation of the aligend sequence. The exon boundaries were identified by aliging protein sequences from known transposases against the consensus *CACTA* DNA sequences found in *B. distachyon* using dot–plot. Boxes below the dotplot indicate the annotated exons for the transposase and ORF2, respectively. The number in the boxes indicate the corresponding exon. Black triangles depict the TIRs. (**A**) *CACTA_A* represents the families where the transposase has 1 and the ORF2 4 exons. (**B**) *CACTA_G* represents the families where the transposase has 2 and the ORF2 5 exons. (**C**) *CACTA_I* represents the families where the transposase has 3 and the ORF2 8 exons.

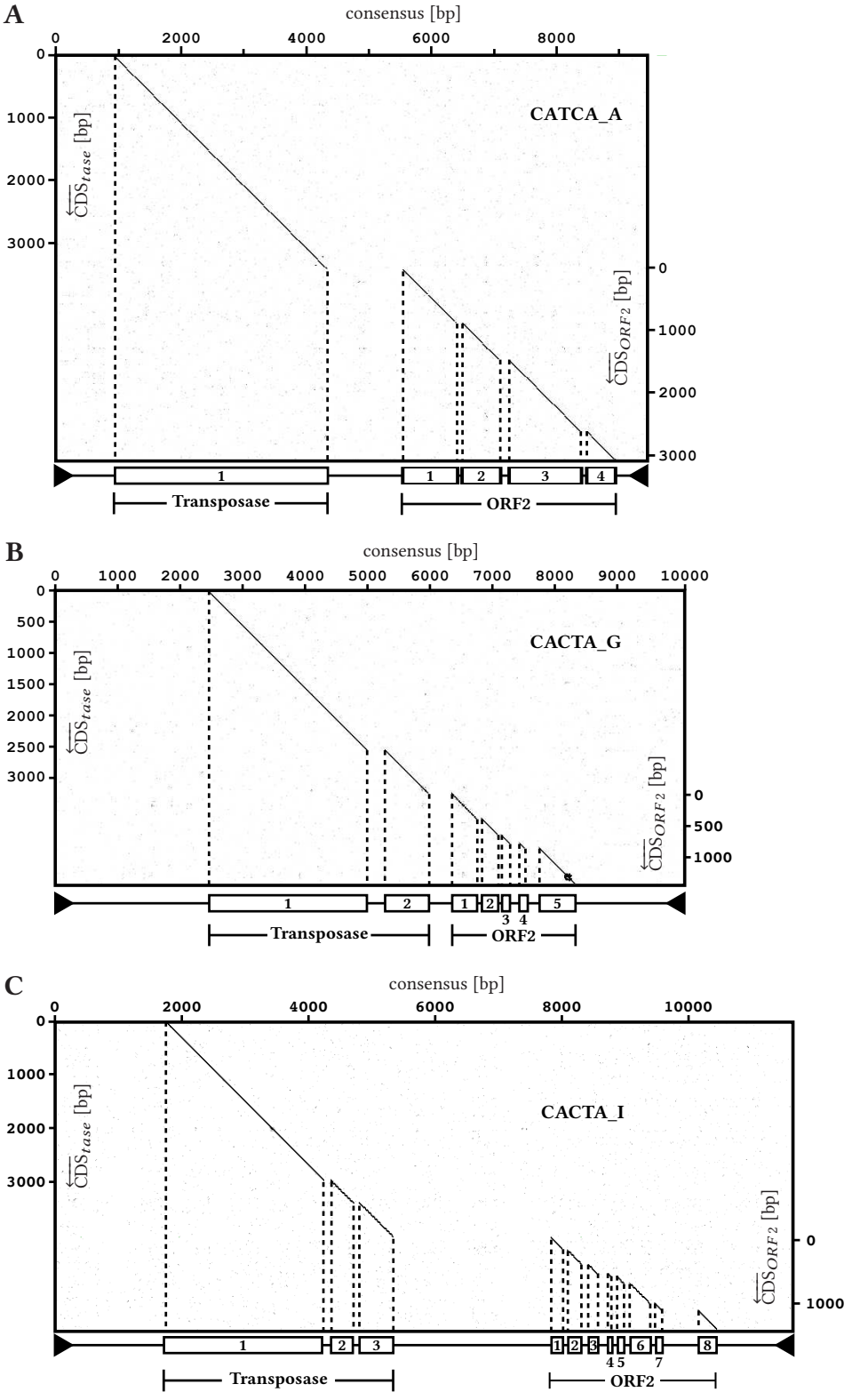

A
consensus [bp]
CATCA_A
Transposase
ORF2
B
consensus [bp]
CACTA_G
Transposase
ORF2
C
consensus [bp]
CACTA_I
Transposase
ORF2

Table 2.2 *CACTA* transposases used in the study.
The number in parentheses after the species name indicates the number of used transposases from the corresponding species. id: name of *CACTA* element; ex: number of exons in the transposase; $^{3'}$: indicates that the transposase is at the 3' end of the element.

A. thaliana (1)		*P. hybrida* (1)		*B. distachyon* (10)		*Z. mays* (9)		*T. aestivum* (10)		*O. sativa* (13)	
id	ex	id	ex	id	ex	id	ex	id	ex	id	ex
Korbin	5	*Psl*	1	*CACTA_A*	1	*Helio*	1	*Caspar*	1	*Calvin*	1
				CACTA_B	1	*Oswald*	1	*Clifford*	1	*Eric*	1
				CACTA_D	1	*Ivan*$^{3'}$	1	*Conan*	1	*Dorian*	1
				CACTA_E	1	*Joey*	2	*Balduin*	3	*Grover*	1
				CACTA_G	2	*Norman*	2	*Horace*	3	*Janus*	2
				CACTA_F	2	*Alfred*	2	*Isaac*$^{3'}$	3	*Baldur*	3
				CACTA_H	2	*En1*	2	*Dario*	4	*Isidor*$^{3'}$	3
				CACTA_I	3	*Preston*	3	*Aron*	4	*Radon*$^{3'}$	3
				CACTA_J	3	*DOPPIA*	3	*Baron*	5	*Rufus*$^{3'}$	3
				CACTA_K	3			*Chester*	5	*Sandro*$^{3'}$	3
										Storm	3
										Sherman	3
										Seamus	4

Maximum-Parsimony and Maximum-Likelihood Methods Reveal Similar Phylogenetic Tree Topologies

To investigate the relationship between the different families, we constructed a phylogenetic tree with the 44 transposase protein sequences. To create a robust tree, we constructed two trees using two different methods and compared them. The first tree was constructed using the program `tree-puzzle` (Schmidt *et al.*, 2002) using maximum-likelihood (ML) and quartet puzzling while the second was done using the program `protpars` from the `PHYLIP` package (Felsenstein, 2005) which is based on maximum-parsymony (MP).

`tree-puzzle` reconstructs trees using a ML method based on quartet puzzling (Schmidt *et al.*, 2002). This method constructs and compares all possible quartets for a sequence dataset to find the most likely relationship for each quartet. Therefore, for N sequences $\binom{N}{4}$ possible quartets have to be analyzed. The analyzed dataset consists of 44 sequences, hence $\binom{44}{4} = \frac{N!}{4!(N-4)!} = 135,751$ possible quartets were analyzed. The tree reconstruction was done in four steps: (i) the model parameters are estimated (pairwise distance matrix and the resulting neighbor joining tree). (ii) ML analysis, where the likelihoods for each quartet are analyzed. Since every quartet has three possible topologies, there are three likelihoods per quartet. Therefore, $3 \times \binom{N}{4}$ topologies are evaluated (407,253 in our case) and the highest likelihood is stored. (iii) The puzzling step creates several intermediate trees. Those trees are constructed by adding random taxa in the branches of the supported quartet topologies where the least contradiction with the relevant quartet trees exists. We used 20,000 intermediate trees for our analysis. (iv) The last step is the consensus step. The intermediate trees are summarized by a majority rule consensus tree where the percent occurrence for each branch is given as puzzle support value. This number can be treated similar to the bootstrap value of parsimony trees. We

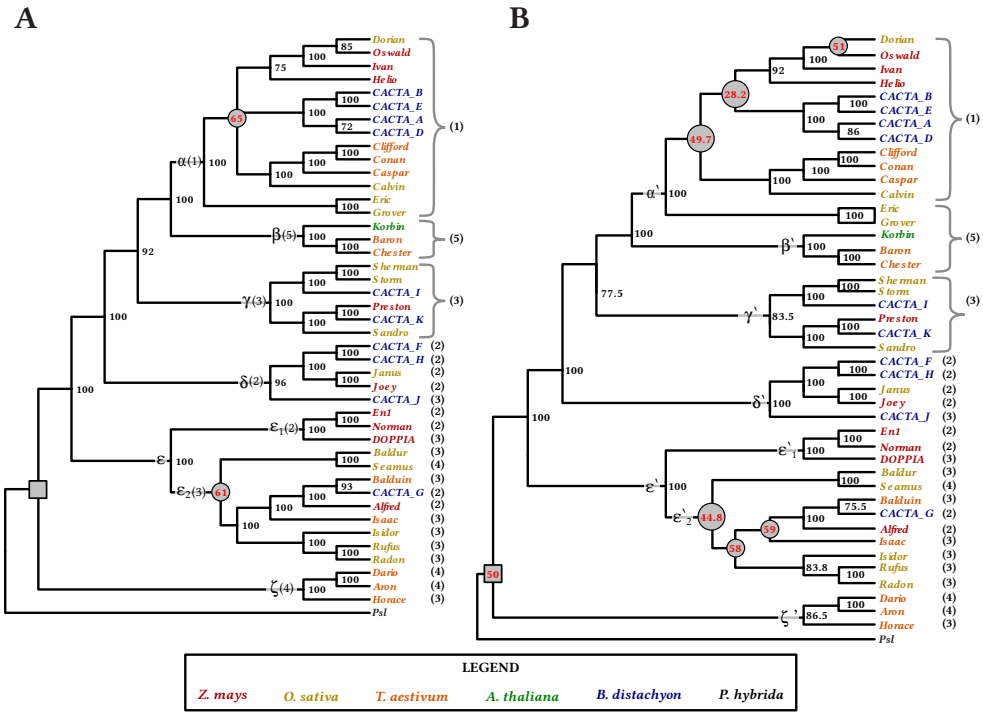


Figure 2.4 The phylogenetic trees derived from transposase protein sequences using ML (**A**) and MP (**B**).
The color of the *CACTA* name indicates its host genome as depicted in the legend. The numbers in parentheses indicate the number of exons per transposase. Gray curly braces indicate clades with the same number of exons. α to ε indicate the clades in the maximum-likelihood (**A**) tree while α' to ε' indicate the corresponding clades in the maximum parsimony tree (**B**). Numbers in parentheses besides the clades indicate the consensus exon number, i.e. the majority of exon numbers per transposase identified for the corresponding clade (for details see text). Gray circles indicate weak bootstrap (<70) or puzzle solving values (<70%), respectively. Gray boxes indicate the root of the tree.

defined the cutoff for weak puzzle support values at 70, i.e. the consensus tree was found in less than 70% of the puzzling trees, in this case 20,000 (Figure 2.4A).

`protpars` from the `PHYLIP` package uses a different approach. It assumes that the change at different sites and lineages is independent and that synonymous base changes are more likely than non-synonymous as well as that the most likely solution is the one with the least steps involved in amino acid changes leading to the difference at one site (Felsenstein, 2005). To prepare the dataset, the sequences

are bootstrapped to generate multiple datasets for the final analysis, in our case 100. To increase the strength of an analysis the order of the input sequences can be changed ("jumbled") for each dataset. We choose to jumble each dataset three times. The resulting trees were compared and a consensus tree using the majority rule was generated. The number how often a certain branch was found in the tree is indicated as bootstrap value. We defined the cutoff for weak bootstrap values at 70, i.e. all branches which appeared in less than 70 out of 100 datasets were considered weak (Figure 2.4B).

The Exon Number in *CACTA* Transposases is not Specific for a Host Genome

The two resulting trees show the same topology (Figure 2.4). We defined six major clades, α to ζ in the `tree-puzzle` tree while the corresponding clades in the `PHYLIP` tree were named α' to ζ'. The clade ε has been subdivided into ε_1 and ε_2. Both trees supported the clades α, β and γ for the transposases with only 1, 5 or 3 exons, respectively. While transposases with only 1 and 5 exons are found in unique clades, the 3 exon transposases are not limited to the γ clade, but found also in the clades δ-ζ. The α clade has a low puzzle solving value (65) and is not fully resolved in the `tree-puzzle` derived tree. The non fully resolved clade contains only *B. distachyon* elements. In the tree derived from `PHYLIP`, the α' clade is fully resolved but has three low bootstrap values 28.2, 49.7 and 51, respectively. However, in both trees the α clades are clearly separated from the remaining clades. The gray boxes in Figure 2.4 indicate the *P. hybrida* outgroup. No puzzle solving value was found for *psl* with the `tree-puzzle` method as the program could not find a suitable root and left the tree unrooted. The tree derived from `PHYLIP` managed to root the outgroup, however with a low bootstrap value of 50.

The `tree-puzzle` approach returned more robust branching values but was

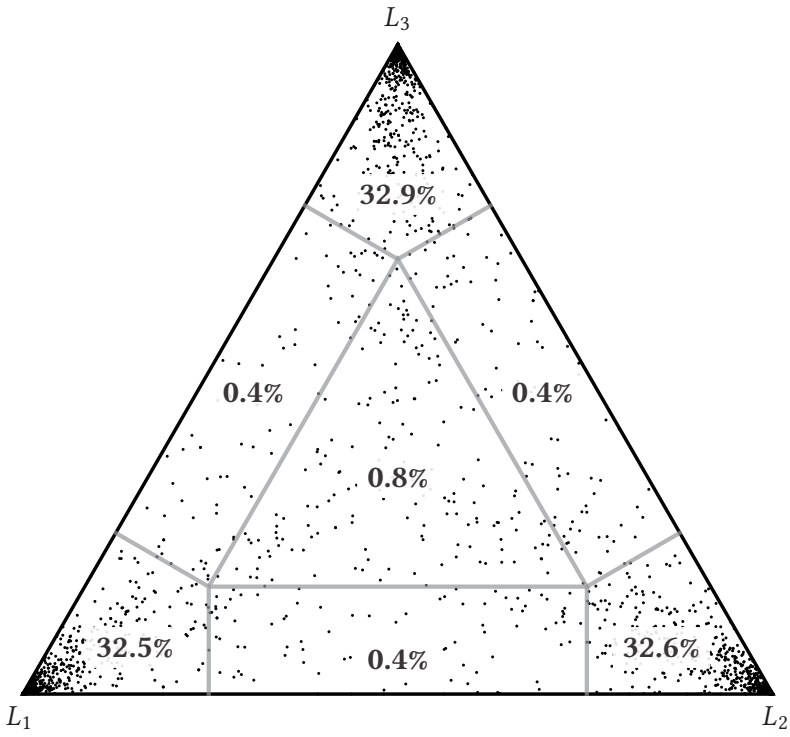


Figure 2.5 Likelihood mapping for the ML tree.
Each black dot indicates the three posterior probabilities (L_1 - L_3) calculated for the 135,751 evaluated tree quartets. The corners represent the number of fully resolved quartets, here 98% which indicates a very robust phylogenetic tree. The rectangles on the side represent the percentage of partially resolved (network-like) quartets in the analyzed dataset. The small triangle in the middle represents the percentage of unresolved (star-like) quartets in the analyzed dataset.

not able to root the α subtree which contained transposases with 1 exon from *B. distachyon*. In contrast, `PHYLIP` rooted all branches but showed lower branching values. Overall, the `tree-puzzle` derived tree has only two low puzzle solving values (<70%) but 2 unrooted branches, one in the α clades while the other is the *Psl*. In contrast, the `PHYLIP` derived tree has no unresolved branches but seven low bootstrap values (<70, Figure 2.4). Nevertheless, both methods returned the same tree topology which supports the overall clustering of the transposases.

To further investigate the evolutionary information in the ML tree, we performed a likelihood analysis (Figure 2.5). This analyzes how many tree quartets

could be totally or partially resolved, indicating the quality, or tree-likeness, of our dataset. In a fully resolved tree all quartets have been resolved, representing a good and robust tree. Trees with partly resolved quartets show network-like structures. Unresolved quartets show a star-like structure, resulting in trees which are bad and have low information value. For robust phylogenetic analysis a fully, or almost fully, resolved tree is needed. Schmidt and von Haeseler (2009) state that, from a biological standpoint, data where 20%–30% are found in a star or network-like structure is not reliable for phylogenetic analysis. The three corners of the triangle represent the likelihood for a specific tree topology from each analyzed tree quartet. The three likelihoods for the three tree topologies for each possible quartet are represented as dot in the triangle. Our analysis showed that 98% of our data is found in the corners, indicating fully resolved quartets. The three 0.4% regions indicate that in total 1.2% of the quartets revealed a network-like structure while 0.8% of the quartets show star-like signals and could not be resolved.

Ancient Lineages Diverged Before the Divergence of Monocotyledons and Dicotyledons

The clade ζ and the subclade ε_1 cluster three transposases each from only one species, maize and wheat, respectively. Otherwise, no species specific clade was detected, e.g. *Korbin* form *A. thaliana* grouped together with *Baron* and *Chester* from wheat in the β clade while the α clades contains 14 transposases with 1 exon from *CACTA* elements out of four different host genomes. This indicates that no evolutionary connection related to the host of the *CACTA* elements exists, i.e. the main branches diverged already before the divergence of monocotyledons and dicotyledons. A closer look at the different clades reveals that transposases with the same number of exons tend to group together, e.g. the clade δ contains five elements of which four have transposases with 2 exons while only one (*CACTA_J*) has 3

exons. The subclade ε_2 contains nine transposases from four different host species of which six have three exons. In addition, four (*Isaac, Isidor, Radon, Rufus*) of six elements found in clade ε_2 harbor the transposase in the 3′ half of the elements. The transposases from the elements *Sandro* and *Ivan* were found in the α and γ clade, respectively. However, should the clades harbor most of the elements with the same exon number, we would expect that all three exons transposases are found in clade γ, which is not the case, which suggests independent exon loss in different lineages. To further investigate the relationships between the transposases, we decided to analyze in detail the exon/intron boundaries.

CACTA Transposase Share Different Intron/Exon Boundaries

To construct a model for the putative evolution of the exon/intron arrangement in transposases, we compared the exon/intron boundaries between the 29 transposases which have more than one exon (i.e. all transposases outside clade α). We defined the exon/intron boundary as the position on the transposase protein where an intron is expected to start (i.e. the consensus splice site `GTAG` on the DNA sequence).

We numbered the intron positions in the 5′→3′ direction, e.g. the first intron of *Baron* is at position 523 while the last one is at position 865 (Figure 2.6, Table A.1). To analyze those positions, we aligned all transposase protein sequences with each other using dot–plot. Since the intron positions indicate the beginning of an intron, we designated the corresponding intron boundary as intron and the number as subscript, e.g. the third intron boundary of *Baron* would be intron_3 (Figure 2.6). Each intron position is represented through a coordinate in the dot–plot. As an example see Figure 2.6 where intron_1 in the *CACTA_K* transposase is at position 745, having the coordinate x = 745 and intron_2 of *Baron* transposase is at position 732, having the coordinate y = 732. This introns create the coordinate pair (745/732) in the dot–plot. This allowed us to check if an intron annotated on one transposase e.g.

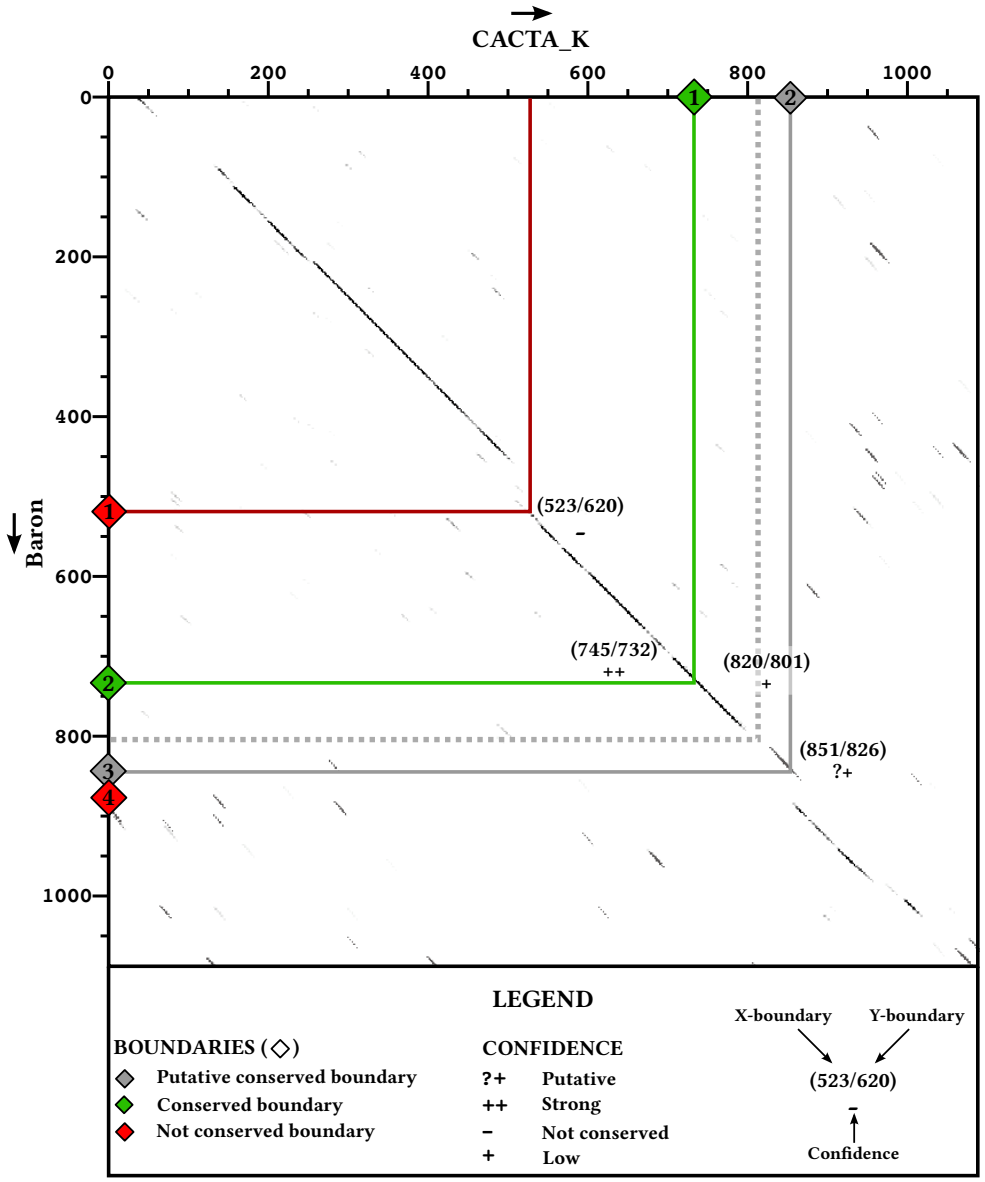


Figure 2.6 Example for identification of conserved exon/intron boundaries in *CACTA* transposases by aligning the corresponding protein sequences using dot–plot.
The transposase protein sequence of *CACTA_K* is depicted on the x-axis while the protein sequence of *Baron* on the y-axis with arrows over the names indicating the orientation of the sequences. Diamonds represent exon/intron boundaries. The color describing the type of boundary as described in the legend. Numbers in the diamonds show the number of the boundary. Colored lines connect corresponding boundaries, with the same color as the boundaries. Numbers besides the connecting lines indicate the position of the boundary in the alignment with the corresponding confidence as described in the legend. An example for a low (+) confidence is shown with the gray dashed line, but has not been identified between *CACTA_K* and *Baron*.

CACTA_K, is conserved on the other transposase e.g. *Baron*. In the example, intron$_1$ of *CACTA_K* (745) and intron$_2$ of *Baron* are found at the coordinates (745/732) in the dot–plot which lies on a conserved stretch in the protein alignment, therefore classified as a conserved intron.

Because transposases are quite divergent in some regions, they can often not be aligned perfectly across the whole protein (Figure 2.6). The alignments showed an overall strong conservation across the aligned transposase proteins. However, small stretches without any homology were found and each coordinate pair of intron positions was classified in one of four levels describing its confidence of being conserved: high (++), low (+), not conserved (−) and putative (?+).

High confidence (++) boundaries were found when the coordinate pair from two intron positions has been identified on a conserved stretch of the two protein sequences (intron$_1$ of *CACTA_K* and intron$_2$ of *Baron* in Figure 2.6).

Low confidence (+) boundaries were found when the coordinate pair from two intron positions has been identified on a diagonal between two conserved stretches of the two protein sequences. An example of a weak signal is given in Figure 2.6, indicated by the gray dashed line. However, *CACTA_K* nor *Baron* have intron positions at those coordinates.

Putative conserved (?+) boundaries were found when the coordinate pair from two intron positions has been identified at borders of diagonals depicting conserved sequence and where found at sites of insertions and/or deletions (intron$_2$ of *CACTA_K* and intron$_3$ of *Baron* in Figure 2.6).

Not conserved (−) boundaries were found when one intron position on a transposase did not correspond to an intron position in the other transposase, even if their intersection was found on a conserved stretch (intron$_1$ in *Baron* in Figure 2.6).

We compared 479 intron positions in total, of which 83 showed strong, 158

weak and 238 putative signals (Table A.2). Comparison of the phylogenetic trees and the table shows high consistency for the five clades as most show a similar pattern of signals when compared to the other clades. Based on the majority of exons per clade, we constructed a loose consensus for the number of exons for each clade. For example, in clade δ with five transposases, 4 transposases had 2 exons while 1 transposase had 3 exons. Since the majority of transposases had 2 exons, we assumed that a representative transposase from clade δ has 2 exons with 1 consensus exon/intron boundary.

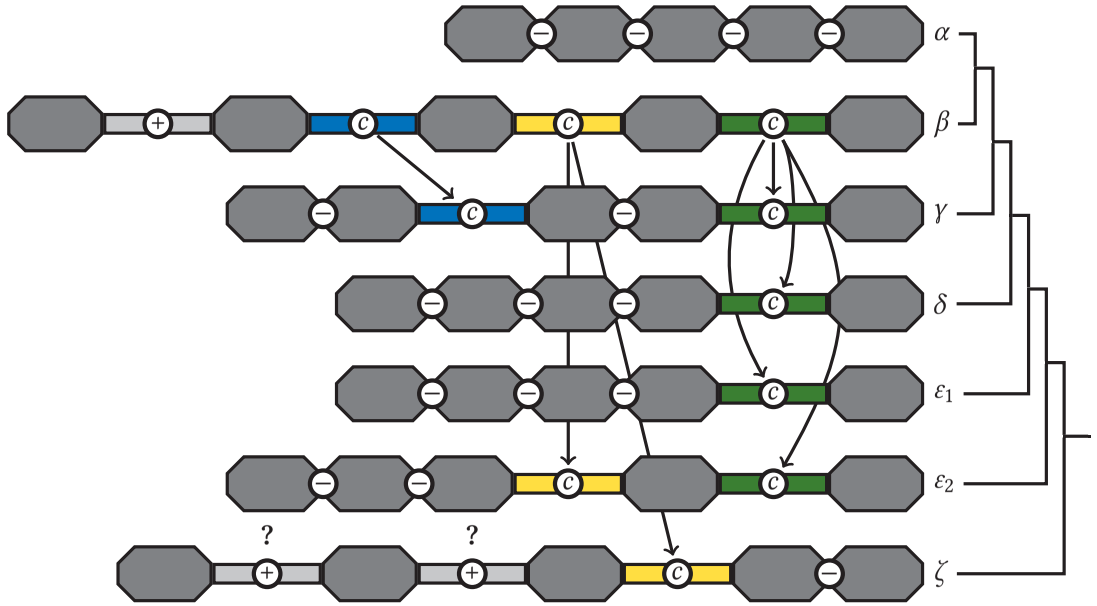
We used this approach for each clade. The representative transposases for clade β has 5 exons. The clades γ and ε_2 have each a representative transposase with 3 exons. The clades δ and ε_1 have both a representative transposase with 2 exons. The clade ζ had too diverse signals, not allowing us to create a clear representative transposase. However, the last intron seemed to be conserved in all members of the clade ζ, creating a representative transposase with three introns, where the first two introns are putative. This simplification allowed us to compare exon/intron boundaries from each clade.

A Model for the Exon/Intron Configurations Found in *CACTA* Transposases

We compared all consensus boundaries to the clade β as it had the largest number of exons (Figure 2.7). The three last introns in the β clade (introns$_{2-4}$) are conserved within the other clades (except α). Intron$_2$ in the β clade is only conserved in clade γ as intron$_1$ while intron$_3$ from the β clade is conserved in the clades ε_2 (intron$_1$) and ζ (intron$_3$). The most conserved intron was intron$_4$ from clade β as it is found within the clades γ (intron$_2$), δ (intron$_1$), ε_1 (intron$_1$) and ε_2 (intron$_2$ Figure 2.7). Except from the intron$_1$ in the clade β and introns$_{1-2}$ in clade ζ, all introns are conserved in one or another clade, e.g. the intron$_4$ from the clade β is conserved

in clades β, γ, δ and ε. This indicates that *CACTA* transposase genes were mostly loosing introns rather than gaining them. However, intron_1 in the clade β as well as introns_{1-2} in clade ζ could be examples of intron gain (Figure 2.7).

Figure 2.7 The model describing the model for loss and gain of introns in *CACTA* transposases.
α to ζ depict the putative transposase exon structure for the corresponding clade. Gray boxes represent exons and colored lines introns. Introns with the same color are shared between clades, also depicted with arrows and ©. Intron loss is depicted with ⊖ while gain of introns is depicted with ⊕. Gray introns are not shared and unique, whereby the ? denote the weak intron signals in the clade ζ. The phylogenetic tree represents a simplified phylogenetic tree from Figure 2.4.



We propose that the transposases in clade β represent the closest exon/intron configuration of an ancient transposase with at least four exons (Figure 2.7). The loss of intron$_3$ in the ancient configuration (β) resulted in the transposases found in the clade γ while loss of intron$_2$ gave rise to the configuration found in clade ε_2. The exon/intron configuration of the clades δ and ε_1 resulted from a loss of introns$_{2,3}$ in clade β. Transposases from clade ζ are derived through the loss of intron$_4$ in the ancient transposase.

The most striking observation was that clade α has lost all of its intronic sequences. The phylogenetic analysis proposes that clade α is very closely related to clade β which has four introns and probably represents the ancient exon configuration. This suggests that the introns in clade α were not arbitrarily lost over time but rather in one single event. This event removed all intronic sequences from clade β, leading to the single exon found in the transposases from clade α.

2–3 Discussion

This chapter described the results of the *CACTA* identification and characterization in the *Brachypodium distachyon* annotation project (IBI, 2010) as well as more detailed findings and analysis of the *CACTA* elements which can be classified in three groups based on the number of exons in the transposases. To further investigate the exon/intron configuration of *CACTA* transposases, we expanded our dataset with 34 additional *CACTA* elements from *Arabidopsis thaliana*, *Zea mays*, *Oryza sativa*, *Triticum aestivum* and *Petunia hybrida* and performed a phylogenetic analysis. In addition, we compared the exon/intron boundaries between the 29 transposases which had more than one exon. This allowed a closer look into the mechanisms of the evolution of transposases in *CACTA* elements.

Our analysis was done with an enhanced *CACTA* database. This explains the higher number of annotated *CACTA* elements, 1,998 compared to the 1,523 were described in IBI (2010) and also explains the higher contribution of *CACTA* elements to the genome sequence, 8.1 mega base pair (Mbp) (2.9 %) compared to the 5.9 Mbp (2.2 %).

Robust Phylogenetic Data

We reconstructed two phylogenetic trees, one using maximum-parsymony (MP) and one using maximum-likelihood (ML). The resulting trees show the same topology, but differ in the quality of the branching. The ML tree is not completely solved at two nodes: the α clade and at the outgroup. The MP tree however solved all branches but shows seven bootstrap values <70%. In contrast, only two puzzle supporting values <70% have been found in the ML tree. The ML method allows to perform a likelihood mapping analysis which indicates if the analyzed data is suitable for phylogenetic inference. The likelihood mapping (Figure 2.5) showed that 98 % of the quartet trees were completely resolved, indicating that our phylogenetic

analysis is very reliable. However, 1.2% of the trees indicated a network-like and 0.8% a star-like tree structure, explaining the unresolved branches in the clade α and the unresolved clade ζ. In addition, it is known that if sequences are too similar, the phylogenetic programs reach their limits since no clear distinction can be made between the different elements. We assume that this is what we are observing at those positions with a low bootstrap/puzzle solving value.

High *CACTA* Diversity Existed Already in the Ancestor of Monocotyledons and Dicotyledons

The phylogenetic reconstruction clustered the transposases rather according to the number of exons than to the host species. Only clade ζ contains transposases which are only found in *T. aestivum*. All other clades have a mixture of host species. The most diverged species are found in β which harbors transposases from *T. aestivum* and *A. thaliana*, two species which diverged approximately 120–340 million years ago (MYA) (Wolfe *et al.*, 1989). This observation shows that already in the common ancestor of monocotyledons and dicotyledons a high diversity among the transposases existed. This observation is supported by comparison of the exon numbers. The analyzed transposases had exon numbers between 1 and 5 while the ORF2 had between 4 and 8. The range of exons in the transposases is similar to previously published *CACTA* transposons. In rice, Greco *et al.* (2005) reported a transposase with 4 exons in *OsESI1* and *Hipa* while studies in maize also indicate several exons for transposases (Masson *et al.*, 1991; Pereira *et al.*, 1986). The comparison between all intron positions from 29 *CACTA* elements with more than one exon showed that *B. distachyon* elements have common introns with transposases from *CACTA* elements identified in other species. In our dataset, a small and very distinctive group of transposases with 5 exons has been identified. All transposases with one exon were found in clade α. This clade consists of

14 transposases from four different species, showing no species specificity for transposases with 1 exon.

The Ancestor Transposase Likely Had Four Exons

The number of exons in the transposases varies between species and our analysis of intron/exon boundaries between the transposases showed that approximately 52 % of the exon/intron boundaries are conserved between two or more transposases. The question arises now if the ancestor transposase contained one exon and gained additional exons or if it contained several exons and lost them over time. In addition, it could be a mixture of both mechanisms where exons are arbitrarily added and lost. In most transposases (20) we annotated exon numbers ranging from 2 up to 4. The conservation of the intron$_4$ from the clade β across several other clades indicates a loss of introns in *CACTA* transposases rather than a gain. However, the unique intron$_1$ of the clade β could indicate that a gain of introns can occur, but is less frequent. Therefore, we assume that the ancestor *CACTA* transposase contained at least 4 exons which afterwards got differentially lost in the different clades and propose that intronic loss is a major force in *CACTA* transposase gene evolution.

Retroposition as Putative Mechanism for Intron Removal

The transposases in clade α, i.e. transposases with 1 exon, could be the final result of intron loss in transposases. However, it is most closely related to transposases with 5 exons. Two possibilities can explain this observation. The transposases with 1 exon are either the final result of a gradual and independent loss of all four introns or a single event removed all introns. A possible mechanism for the loss of introns is retroposition. In this process RNA, is converted into DNA by reverse transcription (Weiner *et al.*, 1986). This could explain the loss of all introns in the transposases of clade α. Retroposition could act either on a transcript of a *CACTA* transposase or on a copy of a *CACTA* element (with introns) which was residing besides an active

retrotransposon and its transcript did not stop at the end, included the copy of the *CACTA* element. Both possibilities lead to a transcribed mRNA which is processed and reverse transcribed, producing a cDNA sequence which gets inserted in a new location.

However, three *CACTA* elements found in *B. distachyon* from clade α have still an ORF2 with 4 exons (*CACTA_A, CACTA_B, CACTA_D*, Table 2.1). This would mean that the introns have been only removed from the transposase but not from ORF2. Thus, simple retroposition does not explain the loss of all introns of a *CACTA* transposase.

Another possibility is that the combination of a transposase with 1 exon and the presence of an ORF2 with several exons is actually a hybrid of two *CACTA* elements. The transposase from an *CACTA* element underwent retroposition and integrated as sole transposase with one exon in a new location. The corresponding ORF2 has been inserted afterwards through another mechanism, e.g. exon shuffling. Exon shuffling describes the creation of new genes through combination of exons from unrelated genes. It has been shown in plants that exons can be added to already existing genes, creating new functions, e.g. cytochrome c1 in potato which acquired its mitochondrial targeting domain through exon shuffling from the *gapdh* gene (Long *et al.*, 1996). This would mean that the *CACTA* elements in clade α are a result of two mechanisms, where the first removed all introns from the transposase and the second added the ORF2 through exon shuffling.

An alternative to the exon shuffling represents the mechanism of gene conversion. Gene conversions constitute a form of homologous recombination which mediates the transfer of homologous genetic sequences (Slightom *et al.* 1980 and reviewed in Chen *et al.* 2007). The initiation of gene conversion in eukaryotes is a double-strand break (DSB) during meiosis. Gene conversion can occur through repair of the DSB through sequence dependent strand annealing (SDSA), whereby

non homologous sequences can be introduced into the region of the DSB (Szostak *et al.*, 1983; Haber *et al.*, 2004). Other proposed pathways share common initiation with the SDSA pathway, but after invasion of the homologous sequence a double Holliday junction is formed (Haber *et al.*, 2004). Cleavage of the Holiday junction by resolution leads either to gene conversion or cross-over. In contrast, if the Holliday junction is resolved by dissolution, only gene conversion occurs (Wu and Hickson, 2003). With such a mechanism, similar to the exon shuffling model, the *CACTA* transposase would undergo retroposition and be inserted in a new location. Afterwards, a DSB repair would have triggered gene conversion, thereby placing a copy of the transposase in front of an ORF2.

2–4 Methods

Putative *CACTA* elements were extracted using the `Perl` program `cacta.pl`. The algorithm as well as the used strategy are described in more detail in the results section (see page 27). Several small `Perl` programs were written to convert or analyze the data derived from the following programs. For genome wide screenings we used `BLAST` standalone version 2.0 (Altschul *et al.*, 1997). The protein division of the TREP database (PTREP, Wicker *et al.* 2002) was used to identify *CACTA* transposases for exon annotation. Visual sequence alignments for exon annotations and classification of intron/exon conservation were done with `DOTTER` version 3.1 (Sonnhammer and Durbin, 1995). Multiple sequence alignments were done using `clustalw` version 2.1 (Thompson and Gibson, 1994). If not stated otherwise, following parameters were used: `-gapext=0.1` for gap extension penalty and `-gap-open=30` for gap opening penalty.

For maximum likelihood phylogenetic tree analysis and likelihood mapping we used the parallelized version of `tree-puzzle`, version 5.2 (Schmidt *et al.*, 2002). Maximum parsimony phylogenetic tree analysis was done using `PHYLIP`, version 3.69 (Felsenstein, 2005). The `Perl` programs developed for this study can be obtained as `git` repository from Jan P. Buchmann (`jbuchmann@botinst.uzh.ch`) or Dr. Thomas Wicker (`wicker@botinst.uzh.ch`)

3 Interspecies sequence comparison of *Brachypodium* reveals how transposon activity corrodes genome colinearity

Jan P Buchmann[1], Takashi Matsumoto[2], Nils Stein[3], Beat Keller[1] and Thomas Wicker[1]

(2012), The Plant Journal, doi:10.1111/j.1365-313X.2012.05007.x

[1]Institute of Plant Biology, University of Zurich, Zollikerstrasse 107, 8008 Zurich, Switzerland

[2]Plant Genome Research Unit, Division of Genome and Biodiversity Research, National Institute of Agrobiological Sciences, 2-1-2, Kannondai, Tsukuba, Ibaraki 305-8602, Japan

[3]Leibniz-Institute of Plant Genetics and Crop Plant Research (IPK), Corrensstrasse 3, 06466 Gatersleben, Germany

Supplementary tables and figures can be found at http://onlinelibrary.wiley.com/doi/10.1111/j.1365-313X.2012.05007.x/suppinfo

Abstract

Intergenic sequences evolve rapidly in plant genomes through a process known as genomic turnover. To investigate the influence of DNA-Transposons on genomic turnover, we compared 1 Mbp of orthologous genomic sequences from *Brachypodium distachyon* and *Brachypodium sylvaticum*. We found that *B. distachyon* and *B. sylvaticum* diverged approximately 1.7–2.0 million years ago. Of a total of 219 genes identified on the analyzed sequences, 211 were colinear. How-

ever, only 24 transposable elements (TEs) out of a total of 451 were orthologous (i.e. inserted in the common ancestor). We characterize in detail 59 insertions and 60 excisions of DNA-Transposons in one or the other species which altered 17% of the intergenic space. The DNA-Transposon excision sites showed complex and highly diagnostic sequence motifs of double-strand break (DSB) repair. DNA-Transposon excision can lead to extensive deletions of hundreds of base pairs (bps) of flanking sequence if the DSB is repaired by "Single Strand Annealing", or insertions of up to several hundred base pairs (bps) of "filler DNA" if the DSB is repaired by "Synthesis Dependent Strand Annealing". In some cases, DSBs were repaired by a combination of both methods. We present a model for the evolution of intergenic sequences in which repair of DSBs upon DNA-Transposon excision is a major factor in the rapid turnover and erosion of intergenic sequences.

3–1 Introduction

Although genes and gene order are usually well conserved in plant genomes, intergenic regions evolve rapidly. Several studies have shown that repetitive fractions of grass genomes are highly dynamic due to the presence of DNA created by created by transposable element (TE) amplification and removed through deletions (SanMiguel *et al.*, 1998; Devos *et al.*, 2002). This "genomic turnover" results in rapid reshuffling of intergenic sequences within a few million years. The term colinearity is widely used to describe the conserved linear order of genes, but in this study we also apply this term to the intergenic space (i.e. intergenic sequences between colinear genes).

DNA-Transposons, also called Class II transposons, move via a "cut and paste" mechanism (reviewed by Wicker *et al.* 2007a). In grasses, DNA-Transposons are characterized by their Terminal Inverted Repeats (TIRs) which can be highly conserved in certain superfamilies such as *Mariner* elements. The TIR motif presumably acts as recognition site for the transposase which binds the DNA at these sites for excision (Brillet *et al.*, 2007; Lewin, 1997; Sinzelle *et al.*, 2008).

DNA-Transposons form two distinct groups: autonomous and non-autonomous elements. While autonomous elements contain the transposase gene and therefore are able to transpose autonomously, but non-autonomous elements lack this gene and are not able to transpose on their own (Yang *et al.*, 2006). Non-autonomous elements often outnumber autonomous ones by far. A good examples are MITEs (Miniature-Inverted-repeat-Transposable-Elements, Bureau and Wessler 1994b) of the *Stowaway* type. In the genome sequence of *Brachypodium distachyon*, 20,994 *Stowaway* elements and only 50 putative mother elements were identified (IBI, 2010). Feschotte *et al.* (2003) proposed a model whereby the high abundance of several thousand *Stowaway* elements is a result of the transposase activity of a few,

distantly related, autonomous mother elements. In a yeast excision assay, Yang *et al.* (2009) found evidence that MITEs achieve high transposition rates by recruiting transposases of autonomous elements.

For insertion, the transposase cuts the DNA, producing overhangs similar to a restriction enzyme. These overhangs are filled in after insertion, creating a so called target site duplication (TSD). The size of the TSD is specific for different superfamilies (Wicker *et al.*, 2007a). For example, *Stowaway* elements have a 2 base pairs (bps) TSD while *Mutator*s usually have 9 bps. When a DNA-Transposon excises from the genome, a footprint in the form of a duplicated target sequence is expected (e.g. for a *Stowaway* this would be the duplication of the TA target site). Excision of a DNA-Transposon leads to a double-strand break (DSB) in the DNA, which has to be repaired by the host cell. This repair may lead to a change in sequence composition at the excision site, depending on the DSB repair pathway used. The simplest DSB repair mechanism is re-ligation, in which the blunt ends at the break are ligated without leaving a trace of the DSB (Gorbunova and Levy, 1999). In the simple sequence annealing (SSA) model, 3′ overhangs are produced by exonucleases. The exposed 3′ overhangs anneal to each other if a short stretch of homology (one or a few bp) is found between them. During repair synthesis, the residual (non-annealed) ends are removed, leading to a deletion (Gorbunova and Levy, 1999; Agmon *et al.*, 2009). In the sequence dependent strand annealing (SDSA) mechanism, the 3′ overhang invades a double-stranded DNA molecule from elsewhere in the genome, triggering synthesis of a copy of a foreign fragment. This is then used as "filler DNA" to repair the DSB (Hartlerode and Scully, 2009).

Previous studies reported deletions or insertions of a few base pairs of foreign DNA following excision of *Mariner* elements *in vitro* (Yang *et al.*, 2006; Robert and Bessereau, 2007) and in single genes in wheat (Mason-Gamer, 2007). However, no comparative study has been performed so far to investigate the quantitative and

qualitative impact of DNA-Transposon excisions on genomic re-arrangements in plant genomes.

Recently, the genome sequence of the grass *Brachypodium distachyon* was published (IBI, 2010). *B. distachyon* belongs to the Pooideae, like barley (*Hordeum vulgare*), wheat and forage grasses. Its genome is approximately 271 mega base pair (Mbp) in size, divided in five chromosomes. *B. distachyon* and wheat diverged approximately 30 million year ago (MYA), while the common ancestor of *B. distachyon* and wheat diverged from rice approximately 40–54 MYA (IBI, 2010). *Brachypodium sylvaticum* is a close relative of *B. distachyon*, but its genome has not been sequenced. Only four large (≥50 kb) genomic sequences of *B. sylvaticum* are publicly available(Griffiths *et al.*, 2006; Bossolini *et al.*, 2007; Faris *et al.*, 2008; Vu *et al.*, 2010). So far, it is not known when the two species diverged.

In this study, we compared approximately 1 Mbp of genomic sequences between *B. distachyon* and *B. sylvaticum* from five loci. Based on comparison of intergenic sequences, we estimate that the two species diverged 1.7–2.0 MYA. Out of 219 annotated genes, 211 were found in colinear positions. In contrast, of 451 annotated TEs only 24 were conserved at orthologous positions in both species (i.e. were present in the common ancestor). Analysis of 192 polymorphic DNA-Transposons indicated that transposon excisions may lead to extensive deletions and/or insertion of hundreds of base pairs of foreign filler DNA. We propose that the excision of TEs is one of the major forces driving the rapid turnover of intergenic sequences and the breakdown of sequence colinearity.

3–2 Results

Gene Order is Highly Conserved between *B. sylvaticum* and *B. distachyon* but Intergenic Regions Are not

We compared five orthologous genomic regions from *B. sylvaticum* and *B. distachyon* using four publicly available sequences from *B. sylvaticum* as well as the *All* locus sequenced here (Table 3.1). To obtain the sequence of the *All* locus, we randomly selected a positive bacterial artificial chromosome (BAC) clone, identified after hybridization of the *B. sylvaticum* BAC library (Foote *et al.*, 2004) with a putative *Aliin lyase* probe, for sequencing. To expand the sequence, we also sequenced one overlapping neighboring clone. The sequences for the *Ph1* (Griffiths *et al.*, 2006) and *Q* (Faris *et al.*, 2008) regions were unfinished *B. sylvaticum* BAC sequences which we concatenated to working models using the *B. distachyon* genome sequence as a template (Table 3.1). We used BLASTN searches against the *B. distachyon* genome to find the orthologous regions in *B. distachyon*. The *Lr34*, *Q* and *Sdw3* region are located on chromosome 1 of *B. distachyon*, and the *All* and *Ph1* regions are located on chromosomes 2 and 4, respectively (Table 3.1). The size of the sequences ranged from 70 kilo base pair (kb) to 360 kb, totaling 1,034,330 base pair (bp) for all five investigated regions (Table 3.1).

A possible concern in such an analysis could be that paralogous loci are compared. We used BLASTN searches to ensure that all five loci are indeed unique in *B. distachyon* within the fully sequenced genome of *B. distachyon*. The *B. sylvaticum* loci for *Lr34*, *Q* and *Sdw3* were used for studies on micro-colinearity and genetic mapping (Bossolini *et al.*, 2007; Faris *et al.*, 2008; Vu *et al.*, 2010). None of these studies found evidence that the loci were duplicated in *B. sylvaticum*.

For the *Ph1* locus, Southern hybridization of two genes using *Eco*RI and *Hind*III-digested genomic *B. sylvaticum* DNA produced only single bands, indicating that

the locus is unique. Segmental duplications in *B. sylvaticum* showing the exact same band patterns would have happened recently. However, this would have no effect on a comparison of *B. sylvaticum* with *B. distachyon* because the phylogenetic distance between the *B. distachyon* locus and either of the duplicated segments in *B. sylvaticum* would be the same. Finally, BAC hybridization, BAC-fingerprinting and chromosome walking at the *All* locus indicated that the *All* locus is unique in *B. sylvaticum*.

In total, we identified 219 genes in both species on the five orthologous loci: 108 genes in *B. sylvaticum* and 111 genes in *B. distachyon*. For the *B. sylvaticum* genes, we used gene identifiers matching those in *B. distachyon* (e.g. the ortholog of *Bradi1g0001* is *Brasy1g0001*). In *B. distachyon* we re-annotated genes where necessary and identified seven genes which had not been annotated previously. (Table S1). A summary of all annotated genes is given in Table S2. The gene count includes 12 tRNAs and two conserved non-coding sequences (CNSs) that were identified by comparison with rice and sorghum (*Sorghum bicolor*). Additionally, we identified 13 putative micro RNAs: seven in *B. distachyon* and six *B. sylvaticum*.

Of the 219 genes annotated, 211 were found in colinear positions indicating a strong conservation of genes between *B. distachyon* and *B. sylvaticum* (Table S2, and see below). A major disruption of colinearity was found at the *All* locus of *B. sylvaticum*: at least 15 kb at the 3′ end at least 15 kb show no homology to the sequence of *B. distachyon* at this locus. We found two pseudogenes in this region which have their closest homologs on chromosome 1 of *B. distachyon*. This indicates a possible translocation derived from the chromosome 1 homologue of *B. sylvaticum*. Alternatively, this sequence could be the result of a duplication in *B. sylvaticum* which introduced a new copy of these genes to the current location. Recent studies have shown that transposable element (TE) activity may cause such duplications of genes to new locations (Wicker *et al.*, 2010). Additionally, three non-colinear

genes were found on *B. distachyon* sequences and two on *B. sylvaticum* sequences, respectively. Visual comparisons of the loci are given in Figures S1–S4 and the annotations in Tables S8–S17.

B. distachyon and *B. sylvaticum* Diverged Approximately 1.7–2.0 Million Years Ago

We calculated the divergence time between *B. sylvaticum* and *B. distachyon* by comparing conserved intergenic sequences (as described in Wicker *et al.* 2003b) and coding sequences (CDSs) of colinear genes using a substitution rate of 1.3×10^{-8} substitutions per site per year (Ma and Bennetzen, 2004). To reduce the influence of possibly conserved regulatory elements such as promoters or downstream elements of genes in intergenic sequences, 1 kb of upstream and downstream of the annotated CDS of genes (or of tRNAs) were excluded. If the remaining intergenic sequence was at least 3 kb long, it was used for divergence time estimation.We were able to align between 3.3 and 30.6 kb of intergenic sequences for the five loci (Table 3.2). The aligned sequences contained 16 of the conserved (i.e. orthologous) TEs but might possibly contain additional, as yet unidentified TE sequences. TEs and other intergenic regions are likely to be methylated which can cause spontaneous conversions from `C` to `T` at `CG` and `CNG` sites. Therefore, we removed all positions which showed `C` to `T` transitions in `CG` and `CNG` sites from the alignments to avoid over-estimation of divergence times due to DNA methylation. The divergence time for each locus was calculated by adding up the results from each investigated intergenic sequence alignment. The individual calculations for the five loci ranged from 2.2 million year ago (MYA) (locus *Ph1*) to 3.4 MYA (locus *Q*) (Table 3.2). A second estimate of the divergence time was obtained by using the coding sequences of genes. To exclude base positions which may be under selection pressure, we used only synonymous sites (see methods). The resulting divergence times ranged

between 2.0 MYA (locus *Ph1*) and 3.6 MYA (locus *Lr34*). In all cases except the *Q* locus, divergence time estimates from intergenic and coding sequences were similar ((Table 3.2)). For the loci *Lr34*, *Ph1* and *Sdw3* both estimates were within each others standard deviation. The discrepancy in the dating results from CDS and intergenic sequences for the *Q* locus may be due to the fact that only relatively few intergenic regions could be aligned. The variability in the divergence time calculations between the loci could indicate the presence of different haplotypes, as described for barley and wheat (Isidore *et al.*, 2005; Scherrer *et al.*, 2005; Wicker *et al.*, 2009a).

To estimate the minimal divergence time between *B. distachyon* and *B. sylvaticum*, we dated the insertion time of all 21 complete and non-orthologous LTR-retrotransposons (i.e. elements that were inserted after species divergence). No full-length LTR-retrotransposons were conserved in the two species (i.e. inserted in the common ancestor). All LTR-retrotransposons are younger than the divergence time of their respective loci and 19 of the 21 Long Terminal Repeat (LTR) retrotransposons are younger than 1.7 million year (Myr) (Table S3).

Analysis of TEs Indicates a Rapid Intergenic Sequence Turnover

In total, we identified 451 TEs, 220 in *B. distachyon* and 231 in *B. sylvaticum*. We identified 64 Class I retrotransposons, belonging to the superfamilies *Copia* (18 elements), *Gypsy* (14 elements), *LINE* (26 elements) and *SINE* (4 elements). Two LTR elements could not be classified into *Gypsy* or *Copia* and 17 were solo-LTRs. Class II DNA-Transposons were the most abundant elements in all five loci. We identified 328 Class II TEs, the majority of the superfamilies *Mariner* (233 elements), *Harbinger* (37 elements) and *Mutator* (35 elements). All 233 *Mariner* elements are Miniature Inverted Repeat Transposable Elements (MITEs) of the *Stowaway* type. The remaining Class II elements were 13 *CACTA*, eight *hAT* and two *Helitron*

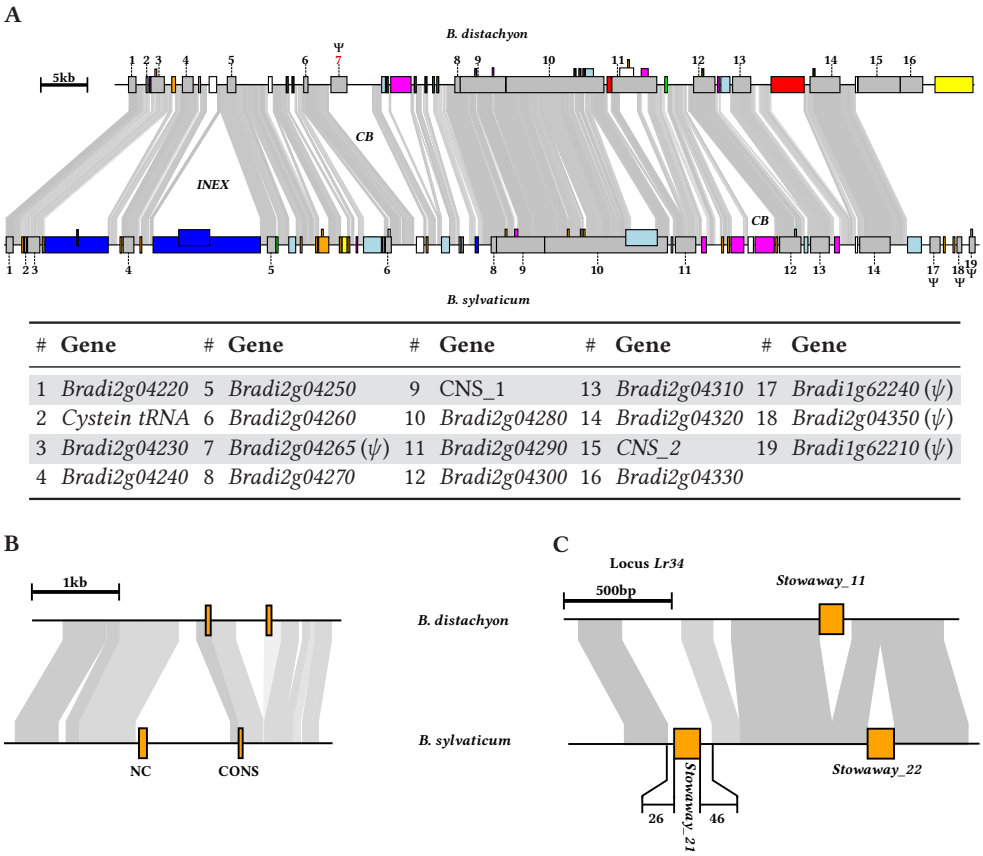
elements.

We also identified 42 TE elements which could not be classified into known superfamilies. They have no homology to known TEs or genes, but BLASTN searches against the *B. distachyon* genome revealed copy numbers between 4 and 100 and they displayed additional characteristics such as target site duplication (TSD) and/or terminal repeats. Most TEs (373) were found in intergenic regions. Only 78 TEs were identified in introns, of which 10 are Class I, 64 Class II and four are unclassified. A summary of all annotated TEs is given in Table S2. Of the 451 identified TEs, only 24 (12 elements in each species) were found in orthologous positions (i.e. were already present in the common ancestor before the divergence of *B. sylvaticum* and *B. distachyon*). One conserved element is a solo LTR, one belongs to the *Mutator* superfamily, while the remaining are *Stowaway* elements.

The Intergenic Regions in *B. distachyon* and *B. sylvaticum* Contain Hundreds of InDels

To analyze the TE-induced changes at the DNA level since the divergence of *B. distachyon* and *B. sylvaticum*, we aligned all loci using the Smith-Waterman algorithm (Rice *et al.*, 2000) where possible, and analyzed the insertions and deletions. Regions that show no homology between the two species and are located between colinear segments (e.g. regions marked ***CB*** in 3.1), could not be aligned. We call such non-alignable regions "colinearity breaks". For all five loci we were able to align 484,723 bp, approximately 47% of the whole analyzed sequence (Table 3.1). The aligned sequences were screened for deletions or insertions (InDels) longer than ≥50 bp (the smallest annotated TE in the *B. distachyon* genome; IBI 2010).

We identified 447 InDels in the aligned sites. The InDels were classified into "TE-related" and "non-TE related", depending on whether we identified TE elements in the InDel sequence. The TE-related set contained 287 sequences, of 18 were excluded



#	Gene	#	Gene	#	Gene	#	Gene	#	Gene
1	*Bradi2g04220*	5	*Bradi2g04250*	9	CNS_1	13	*Bradi2g04310*	17	*Bradi1g62240* (ψ)
2	*Cystein tRNA*	6	*Bradi2g04260*	10	*Bradi2g04280*	14	*Bradi2g04320*	18	*Bradi2g04350* (ψ)
3	*Bradi2g04230*	7	*Bradi2g04265* (ψ)	11	*Bradi2g04290*	15	*CNS_2*	19	*Bradi1g62210* (ψ)
4	*Bradi2g04240*	8	*Bradi2g04270*	12	*Bradi2g04300*	16	*Bradi2g04330*		

Figure 3.1 Visual representation of the comparison of the *All* locus between *B. distachyon* and *B. sylvaticum*.
(A) The annotated TEs are indicated with colored boxes as described in the legend. Nested elements are drawn as elevated boxes. Gray areas connect regions with the sequence similarity described in the legend. Numbers above genes refer to gene annotation. Black gene numbers indicate colinear genes while red numbers depict non-colinear genes. Ψ indicates pseudo genes. ***INEX*** and ***CB*** indicate examples of an insertion/excision of a TE and a break in colinearity, respectively.
(B) Close up of a region with a conserved (CONS) and an non-conserved (NC) DNA-Transposon (*Stowaway*).
(C) Close up of locus *Lr34* showing a region with three *Stowaway* elements. *Stowaway_22* represents am insertion while *Stowaway_11* represents an excision where a sequence fragment on a side of the element was removed. *Stowaway_21* represents a *Stowaway* element found in a region of broken intergenic colinearity with numbers indicating the distance in base pairs to the bordering colinear regions.

due to fragmented TE elements and 11 were excluded due to the presence of multiple TE elements. The problem with multiple TE elements on one InDel sequence is that one cannot clearly determine which, if any, of those elements created the InDel. Hence, the TE related set contained 258 InDel sequences. The non-TE related set contained 160 InDel sequences. We attributed 15 InDels to presence/absence of non-colinear genes (five sites), microsatellites (five sites), tandem duplications (four sites) and one additional intron in *Bsyl1g51250*.

Analysis of DNA-Transposon Polymorphisms

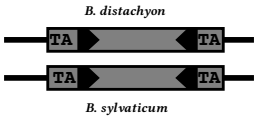
In total, 64 of the 258 TE related InDels were retrotransposon insertions. Retroelements proliferate by a copy-and-paste mechanism, and move "uni-directional" as they can only insert into the genome. In contrast, DNA-Transposon are proliferating through a "cut and paste" mechanisms and move "bi-directionally", i.e. can excise and insert, and both steps leave specific sequence footprints. We focused on the 192 InDels containing DNA-Transposons, distinguishing three cases: conservation of the element, insertion and excision (Figure 3.2 and Table 3.3).

Conserved DNA-Transposons. A total of 11 loci contained conserved (i.e. orthologous) DNA-Transposon. These must have been inserted in the common ances-

Figure 3.2 ***(following page)*** The classifications of the movements of DNA-Transposons found in colinear regions of *B. distachyon* and *B. sylvaticum* using *Stowaway* elements as examples.
The left column shows a representative alignment and the right column a schematic drawing of the corresponding classification. In the alignments, the location of the *Stowaway* elements are indicated with a black line, marking the target site in bold. In the schematic drawings, the *Stowaway* element is depicted as gray box. TA indicates the target site and the black triangles terminal inverted repeats. (**A**) Example of an alignment, where two conserved *Stowaway* elements are found in the same position and must already have been present in the last common ancestor. (**B**) Example of an insertion event, where a *Stowaway* element inserted after species separation. (**C–E**) show excision events which can be further classified as perfect (**C**) or imprecise (**D, E**). (**D**) represents an imprecise excision event where the excised sequence is larger than the excised *Stowaway* element. (**E**) represent an imprecise excision event where we identified some remaining sequence between the target sites.

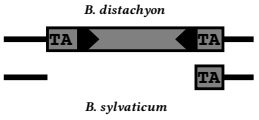
A Conserved element

```
B. distachyon CAACTTTACAATACGCTCATACTCCCTCCATTTCATAAAGG---------
              ||||||||||||||||||||||||||||||||||||||
B. sylvaticum CAACTTTACAATACGCTCATACTCCCTCCATTTCATAAAGGTTGGCGTGT

B. distachyon TTGGTTTCGGTAAGACAAGACTTTCACCATGAATTACTAAATTAATATGT
              |||||||||.|||||||||||||||.||||..|||||||||.|.|.|   |||
B. sylvaticum TTGGTTTCGTTAAGACAAGACTTTGACCAAAAATTACTAATTAA---TGT

B. distachyon GTTTTTTCATACATGAAAT--TTATACCAATAGATTGGTCTTCAAAATTT
              |||||||||||||||||  ||||||||||||||.||||||||||.||
B. sylvaticum GTTTTTTCATACATGAAATATTTATACCAATAGATTCGTCTTCAAAAGTT

B. distachyon CTTGCTAATAATTGTGGTTTCATATCATATAAATTATATTAATTAAGTTA
              |||||||||.|||||  ||||||||||||.||||||||||||||||||
B. sylvaticum CTTGCTAATGATTGT---TTCATATCATATGAATTATATTAATTAAGTTA

B. distachyon TCATTGGTC-AAGACTAGTCTTAACGAAACCAAATACGCTAACCTTTGTG
              ..|||||||  ||||||.|||||||||||||||||||||.||.|||||||
B. sylvaticum GTATTGGTCAAAGACTTGTCTTAACGAAACCAAATACGCCAATCTTTGTG

B. distachyon AACAAGAGGGAGTACTTAATAGGAGCGTTGCCGCTGCCCATGTGAATTAG
              ||.|||||||||||||.||||||||
B. sylvaticum AAAAAGAGGGAGTACTTCATAGGAGC-------------------------
```



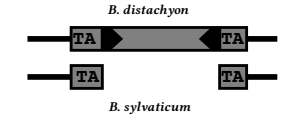
B Insertion

```
B. distachyon ATAGGATTAATTCGGCTCAAATCTAATACTTCCTCAGTTCCTAAATTCGT
              |||||||||||||||||||||||||
B. sylvaticum ATAGGTTTAATTTGGCTTAAATCTAA-------------------------

B. distachyon GTCGTTGTTTTAGTACAAATTTAAACTAAAACAACGACAAGAATTATGGA

B. sylvaticum -------------------------------------------------

B. distachyon ACGGAGGGGAGTAGCAAATTTACTGTCAATCCACCTTAAGAGGCAAAAGCA
                        ||.|||||||.|||||||||||||||||||||||||||.|
B. sylvaticum ----------TAACAAATTCACTGTCAATCCACCTTAAGAGGCAAAAGAA
```



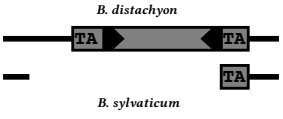
C Perfect excision

```
B. distachyon CATTAGAAACGGAAAC--AACATATTGTCATGCTTAC---CAAGTACTCC
              ||||.||||||||||  |..|.|||||||||||||   .|||||
B. sylvaticum CATTGGAAACGGAAACGTACAACATTGTCATGCTTACATTAAAGTA----

B. distachyon CTCCGTTACATAATTCTTGTCGAAATATTACATATATCTAGACGTTTTTT

B. sylvaticum -------------------------------------------------

B. distachyon AGAAATAAATACATTAATTTTTGGGCAAATTTGAAACAAGAATTATGGAA

B. sylvaticum -------------------------------------------------

B. distachyon CGGAGAAAGTAAATCAAAGCAGAAAGGACAGGGGGAAGCGA--GTTCAGCA
                       |||..|||||||||.||.||||||.|||||  ||||||||
B. sylvaticum ---------TAAGTAAAGCAGAAAAGATAGGGGCAAGCGATCGTTCAGCA
```



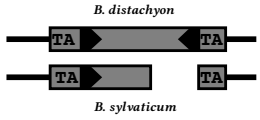
D Imprecise excision with additional sequence removal

```
B. distachyon ATTGTAGTTTGTTTCTTCTTCAAACAATCTTATCACAGAGGGAGAAATTA
              ||||||||||||||.|||||||||||||||||||||||||.||||
B. sylvaticum ATTGTAGTTTGTTTTTTCTTCAAACAATCTTATCACAGAAGGAG------

B. distachyon TACTCCCTCCCTCCCATATTAAATGATTCAAATTTGTCTAAACATGGAAG

B. sylvaticum -------------------------------------------------

B. distachyon TTTCTATATACTAAAATACGTCTAGATACATGTAATATTTCGGCACTTAA

B. sylvaticum -------------------------------------------------

B. distachyon TATAGGACGGAGGGAGTAGCATATTTAATCGTGGTTACTTGTTATTCTTA
                              ||||||||||||||.|||||||||||||||||||
B. sylvaticum ----------------TAGCATATTTAATCATGGTTACTTGTTATTCTTA
```



E Unprecise excision with remaining sequence

```
B. distachyon TGGTGCATCGTAATTGTTCATTTGTACTCCCTCTGTCCCATATTAAACTG
              |||||||||||||||||.|||||||||||||.|||||||.||.|....
B. sylvaticum TGGTGCATCGTAATTGTTAATTTGTACTCCCTCCGTCCCATGTTGACTGT

B. distachyon TACGAAATCAGCGACACTCAATATGGGACGGAGGGAGTAGTATAAGGGAG
              ..|..|.|                              |||||||.|.|||
B. sylvaticum CGCTGATT----------------------------TAGTATACGAGAG
```



tor of *B. distachyon* and *B. sylvaticum* and have not transposed since then. One is a *Mutator* element (*LeChuck*) and 10 are *Stowaway*-MITEs. As both *B. distachyon* and *B. sylvaticum* now contain all 11 elements, there are 22 conserved elements. To classify a TE element as conserved, the entire region (the element and its flanking region) had to be alignable (Figure 3.2a).

Insertion of DNA-Transposons. Another 59 alignments were classified as insertion events: in one of the sequences an element inserted at this position after species separation. The majority were *Stowaway* elements (48 elements), followed by 7 *Mutator*, two *Harbinger* and two *hAT* elements. We classified events as insertions if the alignment showed the TE and its TSD in one sequence while in the other sequence only the nucleotides of one TSD (e.g. TA for *Stowaway*) were present. Thus, the sequence with the insertion contains the TE and the TSD (Figure 3.2b).

DNA-Transposon Excision

We identified 60 alignments containing DNA-Transposons which presumably have excised in one of the two species after species separation. Detailed analysis revealed two classes of excision events: perfect and imprecise excisions. A perfect excision we defined as complete removal of the TE with retention of the two copies of the TSD. We found only one perfect excision, a *Stowaway* element ((Figure 3.2c) and (Figure 3.3a).

Figure 3.3 ***(following page)*** Listing of all 60 identified excision events in which one breakpoint is exactly bordering the TE.
The point of excision plus 10 bp of flanking sequences are depicted. For better visibility the target site is separated from the flanking sequence. Deleted bases are depicted as "-". Gaps in the alignment larger than 10 bp and deletions larger than 40 bp are indicated in parentheses. Filler DNA is underlined. The number of examined cases is given in parentheses. DTT: *Stowaway*, DTM: *Mutator*, DTH: *Harbinger*, DTC: *CACTA*, DTA: *hAT*

A Perfect excision (1)

```
TACATTAAAG  TA                                                  TA  AGTAAAGCAG
```

B Excisions with deletions (13)

DTT

```
TTCATCT---                                                      TA  TAAAAAAGGC
CCTTAAT---                                                      TA  CCTTCCTTTT
GCATTA----                                                      TA  TATAGTTAAT
TGT-------                                                      TA  AATGAAATCA
GGAG------                                                      TA  GCATATTTAA
(390bp del)                                                     TA  CTTTGTACCA
TGAACTATAT  TA                                                      ---AAACTTT
CAATGCAAGG  TA                                                    (20bp del)
GATTAATAAA  TA                                                  (1055bp del)
```

DTM

```
CGTGCACTGC  CCACAGAGCT                                              (42bp del)
(12bp del)                                                  TAAGTTA TTAGCAAACA
```

DTH

```
GCAATCTTAC  TTA                                                     ----GCTAGC
```

DTC

```
(17bp del)                                                          TTCGCCGTCC
```

C Excisions with deletions and filler DNA (44)

DTT

```
GGCCAGCATA  TAG                                                 TA  CTATTTTGGA
GAGGAAAACA  TAC                                                 TA  GCTTTTTTTA
GAAATAGTAG  TACGGAA                                             TA  TTTGCGTTTG
ATTAGTGGTT  TAAT                                                    -----TATCA
TACAAAGTAG  TACTCCACTGTGCC                                          ---------T
CTAATAATGC  TACC                                                    AGTTGATTTA
TGTTAAATTC  TACCAA                                                  (25bp del)
ATAAAAATCT  TACGCC                                                  GCTTACTTTG
GTCAGGATCA  TAGAAAAAA                                               ---------T
GATATAGTAG  TAGTAGTACTGTCTA                                         ---TGTAGCA
TGTTTCTATT  TACTCCTATTTTAGG                                         ----TGTACA
TAAATATTAT  TACTACATCCGATCCGTTCTTAAATATAA                           (11bp del)
GACTCT----                                                  AGTA    TTTCTTGTTT
GATTGT----                                                  TGCACATA TGTAGATATT
(11bp del)                                                  AAAAAATA TCAGAGTTTT
----------                                                  TGGCATGTTA TATGCTGCGT
CATCATATA-                                                  CGTAGTAGTA TCATTTTTTG
----------                                          GTAAATAACTATCTATCTATA ATCCGTTTCA
(13bp del)                                  GCCTTAATAGTATGTACAGCTCATGTTA TTACTTTATC
GGGATTTC--                          ACGAAAACATATGCATATCGTGGCGTGTACCATA TCACGGACAC
(46bp del)          AATTACTACTACGTGCTAAAACTATAGGCTACTCCCACCAGCTA CCTTTCTTAC
TTTGCCCTAG  TA                                                  CCGTATTA CATTCTTAAA
(17bp del)  AAGCAGACG...        (49bp filler)       ...CACGTCAGTA TAATTTTATC
GAGTAAATTA  CAG                                                 AAACAGCCAC
AAAATAGTAA  TCCT                                                    ----AAATGA
CTTAAACTTG  CATGTATCTGATCAAAC                                       (14bp del)
(52bp del)  CTCCGTCTTATACTACGA                                      CAAACTACTG
TCTTGCGATG                  GTTCATGATTAACCCCATGGAAAATAGTACAGTA AGGCTTCCTT
ACGGCAATCG  ATGGACGTAT...       (62bp filler)       ...CTCTCTATCA TGAAAAATTC
```

DTH

```
GTAAGTTTTC                                                  CAAATTTA GCAGTATTCT
(14bp del)                                                  CCTGGTAA CCACGCCCAC
(17bp del)                                                CGTGAAAAAATT GCTGCTCCTG
AGGAG-----                                          CAGACCACGCCCGGCCCCTAA GCCAGCCTTT
TGGGATAGGC  TAAG                                                    -TAAGATCTC
ATCGAAAGAC  TAAAAATT                                                AATTAAAGCC
(34bp del)                                                  GGGCCT  GCAAAACTTA
TTA-------                      TAAAAATTATTTGTCTTTTCCGAACTCTTAACC    TAAGTCCCTA
(2432bp del) ATATATGACC...      (769bp filler)      ...GAGTGCACCT   TAGTTTAGTC
CCAAATACCT   GCAGGATAGG...      (240bp filler)      ...AAACATACCC   AGCCCAAATA
TAATCAAATA   ATATGCAGGT...       (39bp filler)   ...AGGTCAGATCTAA   AACCCTCACA
```

DTM

```
(89bp del)              CTGAGGCATACGCCGGCACGCGGCGCTTGTCTACTGGAACA    AACGCCAATC
GGGCCGACCA                                      TTCGTATTTTTCGTTGTC   (16bp del)
ATAA------                                      AAATTAAATTAAACATAATC GGAGACAGCA
```

DTA

```
GGAGGAC---                                          TTAACATGGGAGGAC  ATGTCTTCGC
```

D Partial removal of elements (2)

```
TGTTAATTTG  TACTCCCTCCGTCCCATGTTGACTGTCGCTGATT                  TA  GTATACGAGA
GAAAGTTATA  TACTCCACATGGG                                       TA  CTGCTTTTAG
```

DNA-Transposon Excision May Cause Deletions of Flanking Regions and Insertions of Filler DNA

For 13 excision events (nine *Stowaway*, two *Mutator*, one *Harbinger* and one *CACTA* element), only one side of the excision was precisely locates at the border of the DNA-Transposon, while the other border was not longer detectable due to additional sequence removal (Figures 3.2d and 3.3b). We postulate that these events represent TE excisions because we consider it highly unlikely that a random deletion would create a breakpoint precisely at the border of a TE (see methods). We observed seven small deletions (3–7 bp), and six deletions that were larger, with sizes ranging from 12 to 1,055 bp (Figure 3.3b). A total of 1,566 bp of non-TE sequences were removed in these 13 events. As excision of an element induces a double-strand break (DSB), we searched for sequence motifs diagnostic for DSB repair. If the DSB is repaired through simple sequence annealing (SSA), one expects the result to be a deletion. After exonuclease digestion, an exposed 3′-overhang triggers the repair upon finding a few base pairs of microhomology on its counterpart (the 3′ overhang of the complementary strand). Therefore, a few base pairs of the sequence just after a breakpoint (inside the non-colinear region) are homologous with the sequence at the other breakpoint (inside the colinear region) or vice versa (Figure 3.4). In the 13 cases where *Stowaway* elements bordered the breaks of colinearity, we identified nine such SSA signatures with sizes between 2 and 6 bp.

Removal of Flanking DNA and Insertion of Filler Sequences after DNA-Transposon Excision

We identified 44 cases where the excision site contained foreign "filler" sequences (Figures 3.2e and 3.3c). The presence of filler DNA at sites which undergone DSB repair indicates DNA repair through sequence dependent strand annealing (SDSA), Puchta 2005). This process, like the SSA, creates 3′ overhangs. In contrast to SSA,

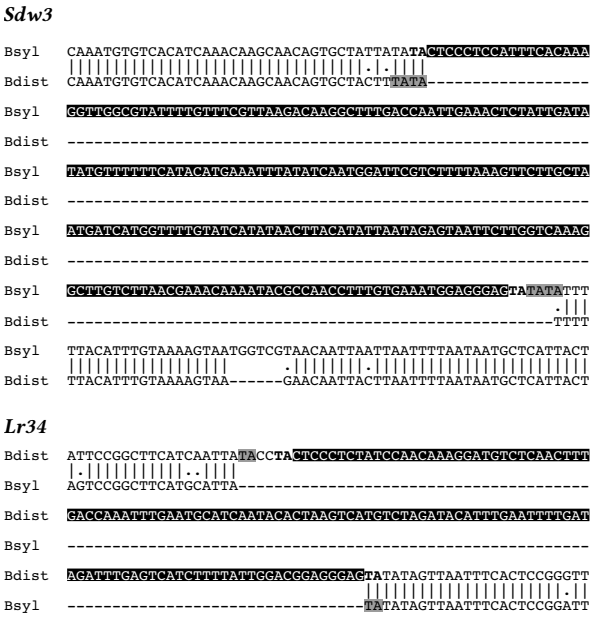


Figure 3.4 Examples of two SSA signatures found at the loci *Sdw3* and *Lr34*. *Stowaway* elements are depicted by the black box while the target site TA in bold. The SSA signatures are indicated by gray boxes.

those overhangs invade a foreign double-stranded donor molecule, triggering repair synthesis. In four cases we detected the TA target sites on both ends and filler DNAs with length of 1 and 6 bp. In 40 alignments, only one target site and insertions between 2 and 789 bp were present. In addition, deletions between 3 and 2,432 bp were also detected in those 44 cases (Figure 3.3c). The described excision events removed 2,993 bp and inserted 1,811 bp of filler DNA. Including the 13 cases for which SSA repair is proposed to have occurred, a total of 4,562 bp were deleted. In all the cases described here, one side of the excision was precisely at the border of the TE (Table 3.3). A summary of all excisions which are precise at least on one side of the TE is given in Table S4.

Imprecise Excision of DNA-Transposon

We found only two cases in which a *Stowaway* element was partially removed and parts of it were still present at the excision site (Figures 3.2e and 3.3d). These

the only examples where a DNA-Transposon was not completely excised. In the example shown in Figure 3.2e, we hypothesize that transposon excision was precise on one side but part of the other terminus remained in the genome. The resulting DSB was repaired by the insertion of a filler segment, explaining the more divergent stretch in the alignment (Figure 3.2e).

Removal and Insertion of DNA on both Sides of DNA-Transposons

Above we only considered excision events were one breakpoint precisely bordered the end of the Class II element, allowing us to classify the event as excision with high certainty. However, we also identified 51 cases in which DNA-Transposon excisions presumably caused deletions on both sides of the element ((Table 3.4), example in Figure 3.1c). In total, 12 alignments (nine *Stowaway*s, two *Harbinger* and 1 *Mutator* element) showed deletions that suggest DSB repair via SSA. We identified six SSA signatures in those alignments: at the *All* locus, one signature showed perfect sequence homology over 8 bp, while two others showed 12 and 7 bp signatures with one and two mismatches, respectively. At the *Lr34* locus, we identified an 11 bp SSA signature. and two alignments, one at the *All* locus and one at the *Sdw3* locus, showed only a microhomology of 1 bp. These six putative SSA repair events removed a total of 2,628 bp of sequence. In the six cases without a SSA signature, between 10 and 2,483 bp of flanking sequences were removed. In all 12 cases, a total of 5,333 bp of flanking sequences were deleted.

In 39 cases (25 *Stowaway*, eight *Harbinger*, three *Mutator*, two *CACTA* and one *hAT* element), a deletion combined with insertion of filler DNA was observed, indicating DSB repair via SDSA (Table 3.4). The filler sequences ranged in size from 1 to 639 bp. Removal of flanking sequences deleted between 1 and 2,371 bp (for example, *Stowaway_21* in Figure 3.1c). These 39 putative DSB repairs inserted 3,278 bp and removed 11,008 bp of flanking sequence.

In total, the 51 putative transposon excisions described here removed 18,969 bp of DNA and inserted 3,278 bp of filler DNA (Table 3.4). A summary of all the elements with deletions on on both sides is given in Table S5. Combined with the 60 excision described above where one end of the DNA-Transposon was precisely bordering the excision site, DNA-Transposon insertion events removed 23,531 bp and inserted 5,089 bp, i.e. a change of 28,620 bp, representing 5.9% of the aligned sequences.

Distributions of Deletions are Tightly Associated with TE Excision

We proposed above that excision of DNA-Transposons sometimes causes extensive deletions of genomic sequences flanking the transposon. This required confirmation, as it could be argued that a deletion caused by something other than the excision could by chance remove an entire transposon plus some of its flanking regions. To test whether the occurrence of such deletions is non-random, we compared the observed data with simulations of randomly distributed deletions. We ran simulations for two loci in *B. sylvaticum* (*All* and *Lr34*). We considered all sequences which are absent in *B. sylvaticum* and are not clearly explained by simple TE insertions in *B. distachyon* (see methods). *B. sylvaticum* contains 26 such deletions at the *All* locus and 92 at the *Lr34* locus. We ran 1000 simulations for each of the two loci to provided a dataset for the expected random distributions of deletions that could be used for statistical testing of the observed data (see methods). We found that deletions that cover entire TEs plus flanking regions are highly over-represented in the observed data at a confidence level far exceeding $p<0.0005$ (Table S6). For example, at the *All* locus, 1.4 of 26 random deletions are predicted to cover entire TEs but in fact, nine of 26 covered entire TEs. Based on the expected values from the 1000 simulation, the probability for such an event to occur by chance is approximately $1.3e^{-12}$. This demonstrates that the distribution of deletions is

non-random.

Deletions Independent from TE Excision

We identified deletions inside the 24 orthologous TEs that most likely occurred independently from TE excisions. In these cases, the TE is conserved in both species but a segment inside the TE is missing in one or the other species, indicating that these deletions were caused by something other than a TE excision. We detected a total of seven deletions in four of the orthologous TE pairs (Table S7). Most are small, ranging from 11 bp in solo-LTR_3 (*Lr34* locus) to 171 in the *Mutator* element *LeChuck* from the *All* locus (Table S7). *LeChuck* also contained the single largest deletion of 3,906 bp, which removed 83% of the element in *B. distachyon*, almost its entire transposase coding region. In total, these seven deletions removed 4,306 bp, or 47% of the total length of the 24 orthologous TEs. DSB repair signatures were found in two deletions.

Table 3.1 Sequences used in *B. sylvaticum* and *B. distachyon*.
GB ID: Genebank ID for *B. sylvaticum* sequence, length: length of analyzed colinear region, aligned: number of aligned bp between *B. sylvaticum* and *B. distachyon* at this locus, %: percentage of aligned bp relative to whole length of analyzed locus, Chr: chromosome in *B. distachyon*, Begin: Begin of locus on indicated chromosome, End: End of locus on indicated chromosome.

Locus	GB ID	length [bp]	aligned [bp]	%	Chr.	Begin [bp]	End [bp]	Reference
Ph1	AM072969	70,427	50,377	72	4	39,099,953	39,214,953	Griffiths *et al.* (2006)
Q	EU153459	76,510	30,140	39	1	2,539,950	2,639,950	Faris *et al.* (2008)
All	HE650836	155,271	77,762	50	2	2,929,243	3,104,514	This study
Sdw3	2768021	360,655	149,659	41	1	18,522,971	18,942,971	Vu *et al.* (2010)
Lr34	FJ436983	371,467	176,785	48	1	49,449,261	49,890,143	Bossolini *et al.* (2007)
Total		1,034,330	484,723	47				

Table 3.2 Summary of the divergence time analysis between *B. sylvaticum* and *B. distachyon* using intergenic sequences.
IG: intergenic, CDS: Coding sequence MYA: million years ago, Sites: total of analyzed base pairs, Div. Time: calculated divergence time, SD: standard deviation.

Locus	IG Sites [bp]	Div. Time [MYA]	SD [MYA]	CDS Sites [bp]	Div. Time [MYA]	SD [MYA]
All	18,069	2.519	0.076	2,768	3,361	0.226
Lr34	30,594	3.376	0.069	6,613	3.580	0.153
Q	3,311	3.426	0.207	885	2.169	0.315
Ph1	4,519	2.257	0.142	1,624	2.011	0.223
Sdw3	19,180	2.565	0.073	5,562	2.561	0.138

Table 3.3 Summary of DNA-Transposon polymorphisms where one border of the excised fragment was precisely at the border of the TE.
fam: superfamily #: number of analyzed elements of this superfamily, ex: number of excisions, subscripts indicating cases with SSA signatures, in: number of insertions, c: number of conserved elements, Δ: removed DNA, fill: inserted filler DNA. DTT: *Stowaway*, DTM: *Mutator*, DTH: *Harbinger*, DTA: *hAT*, DTC: *CACTA*.

Species	fam	#	ex	in	c	Δ[bp]	fill[bp]
B. sylvaticum	*Mariner*	58	26_3	22	10	1,692	269
	Mutator	12	5_1	6	1	183	87
	Harbinger	9	8_1	1	0	93	127
	hAT	2	1	1	0	11	7
	CACTA	1	1	0	0	17	0
	Total	82	41_5	30	11	1,996	490
B. distachyon	*Mariner*	51	15_4	26	10	92	258
	Mutator	2	0	1	1	0	0
	Harbinger	5	4	1	0	2,474	1,063
	hAT	1	0	1	0	0	0
	Total	59	19_4	29	11	2,566	1,321
Total	*Mariner*	109	41_7	48	20	1,784	527
	Mutator	14	5_1	7	2	183	87
	Harbinger	14	12_1	2	0	2,567	1,090
	hAT	3	1	2	0	11	7
	CACTA	1	1	0	0	17	0
	Total	141	60_9	59	22	4,562	1,811

Table 3.4 Summary of putative DNA-Transposon excisions which removed flanking sequences on both sides of the element.
fam: superfamily ex: number of excisions with subscripts indicating cases with SSA signatures, Δ: removed DNA, fill: inserted filler DNA. DTT: *Stowaway*, DTM: *Mutator*, DTH: *Harbinger*, DTA: *hAT*, DTC: *CACTA*

Species	fam	ex	Δ[bp]	fill[bp]
B. sylvaticum	*Mariner*	14_3	9,878	621
	Mutator	3	879	80
	Harbinger	6_1	2,677	1,506
	hAT	1	11	4
	CACTA	2	135	46
	Total	26_4	13,580	2,257
B. distachyon	*Mariner*	20_2	5,217	992
	Mutator	1	4	0
	Harbinger	4	168	29
	Total	25_2	5,389	1,021
Total	*Mariner*	34_5	15,095	1,613
	Mutator	4	883	109
	Harbinger	10_1	2,845	1,531
	hAT	1	11	4
	CACTA	2	135	46
	Total	51_6	18,969	3,278

3–3 Discussion

To study recent events in genome evolution, we compared a total of 1 mega base pair (Mbp) of genomic sequences in the recently diverged species *B. distachyon* and *B. sylvaticum*. We observed strong colinearity of genes (211 of 219 genes were conserved), but the intergenic space has diverged almost completely: only 24 (12 in each species) of 451 transposable elements (TEs) were conserved in orthologous positions.

We used conserved intergenic sequences and coding sequences from colinear genes to estimate the divergence time of the two species. For the individual loci, we estimated divergence times ranging from 2.0–3.4 million years ago (MYA). Divergence time estimates derived from intergenic regions were largely consistent with those derived from synonymous sites in the coding sequences of genes, indicating an overall robustness of the estimates. The variability in divergence times between the loci is possibly due to the presence of different haplotypes. Previous studies have shown that many plant genomes are a mosaic of haplotypes of different ages (Isidore *et al.*, 2005; Scherrer *et al.*, 2005; Wicker *et al.*, 2009a). As haplotypes may be older but cannot be younger than the actual species, the youngest divergence time estimate (2.0 MYA) is probably closest to the actual species divergence. The dating of the Long Terminal Repeat (LTR) insertions showed that all non-orthologous LTRs were younger than the divergence time of the loci which they were inserted, further supporting our divergence time estimates. These data also allowed to further narrow down the estimated divergence time. Except for two, all LTR retrotransposons were younger than approximately 1.7 million years (Myr). Excluding the two outliers, we propose that *B. distachyon* and *B. sylvaticum* diverged at 1.7–2.0 MYA.

SanMiguel *et al.* (2002) proposed that, after 10–14 Myr, any similarity in the intergenic regions between barely and wheat will be gone, making comparative anal-

ysis of intergenic sequences impossible. However, the fact that they did not find LTR retrotransposons older than approximately 2.2 Myr (using the updated substitution rate of Ma and Bennetzen (2004) suggested an even more rapid genomic turnover. Additionally, Ma *et al.* (2004) analyzed the dynamics of 11 LTR-retrotransposon families in rice, and calculated that most insertions happened less than 6 MYA. Furthermore, Hurwitz *et al.* (2010) calculated that approximately two-thirds of the LTR-retrotransposon in different rice species inserted <0.58 MYA. Our study adds a dataset for Brachypodium, a grass with a small genome, showing that intergenic sequences are mostly rearranged after only 1.7–3.4 Myr. This indicates that the pace of intergenic sequence turnover is independent of genome size and TE content.

"Sloppy" DNA Repair Blurs Footprints of Transposon Excisions and can Lead to Major Breaks in Sequence Colinearity

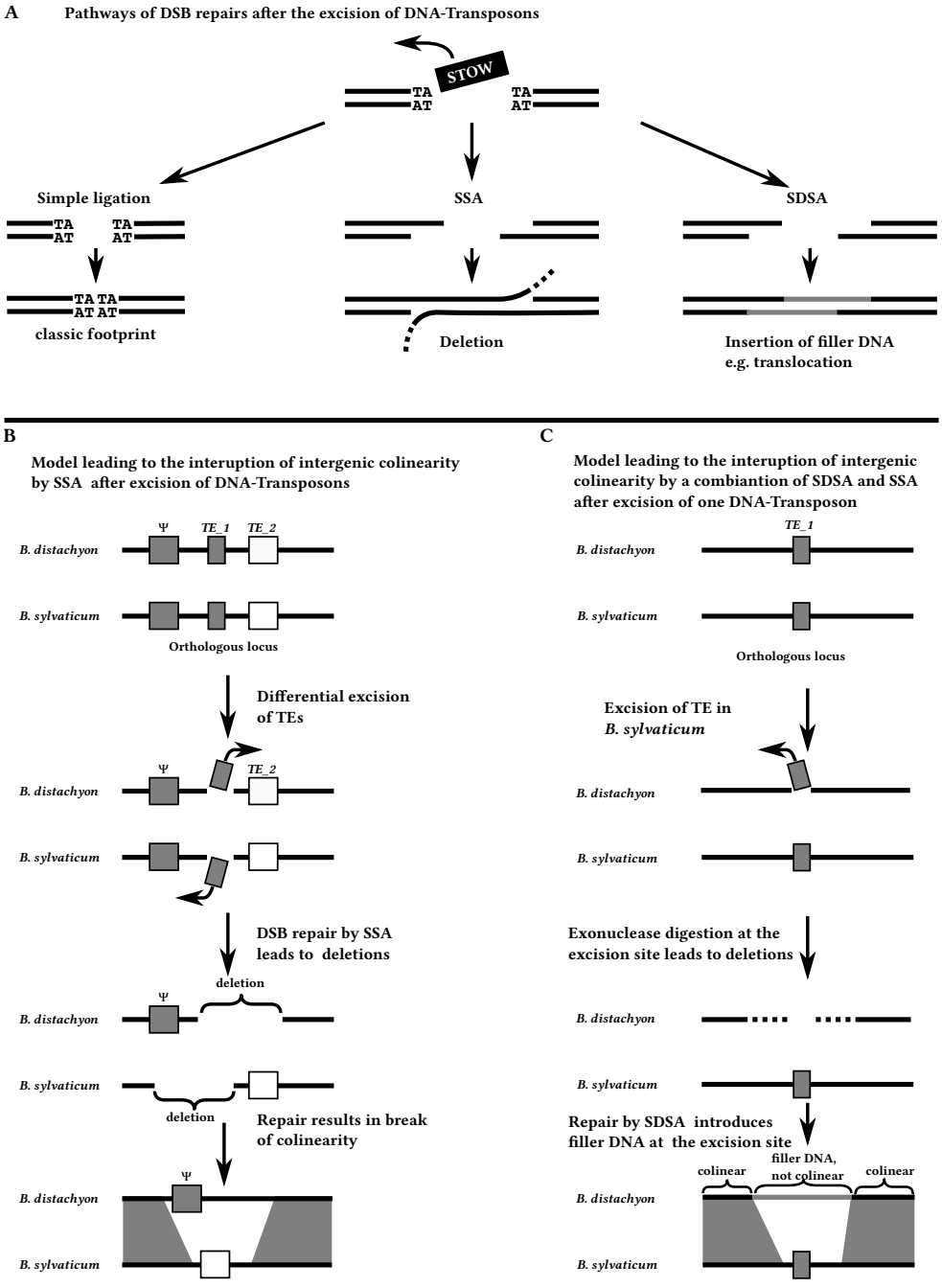
The most interesting insights came from the analysis of DNA-Transposon polymorphisms. Although all 59 insertions showed precisely the expected signature, consisting of the element and its target site duplication (TSD), almost all putative excisions did not. In fact, we identified only one perfect excision in which a *Stowaway* element had excised after species divergence, leaving a `TATA` footprint. It has been reported that the excision of *Ac/Ds* elements, which belong to the *hAT* superfamily, create a 1 base pair (bp) overhang at the excision site if expressed in yeast (Weil and Kunze, 2000). Likewise, excision of a *Stowaway* element creates small 3′ overhangs deriving from the Terminal Inverted Repeats (TIRs) (Dawson and Finnegan, 2003; Yang *et al.*, 2006; Robert and Bessereau, 2007; Richardson *et al.*, 2009). Therefore, assuming that excision of DNA-Transposon generally leaves overhangs at the excision site, a perfect footprint is observed only when the 3′ overhangs are removed by exonucleases and the blunt ends are directly ligated (Figures3.5a and S5).

The finding that most DNA-Transposon excision sites showed a considerable

alteration in sequence composition was intriguing, but is in agreement with previous studies that found that accurate double-strand break (DSB) repair in plants is rare (Gorbunova and Levy, 1999). In 60 cases, it is highly likely that the InDel is due to a transposon excision, because one side precisely borders the excision site. Diagnostic sequence motifs at the borders of the deletions are completely consistent with DSB repair via the SSA pathway. In fact, in the 13 cases in which DNA was was deleted by SSA repair, we identified stretches of microhomology (2–12 bps) precisely at the expected positions flanking the deletions. In the cases where no SSA signatures were found, the borders of the deletions might be degenerated or caused by a different DSB repair mechanism. DSB repair by SSA leads to colinearity breaks if DNA-Transposons that form a pair of neighboring, orthologous elements excise in *B. distachyon* and in *B. sylvaticum* (Figure 3.5b).

In addition to the excisions that caused deletions in flanking sequences, we found 44 cases in which filler DNA was inserted (Figure 3.3c). All but three sites showed a combination of insertion of filler DNA and deletions of the sequences flanking the excised element. We propose that, in most cases, exonuclease activity

Figure 3.5 ***(following page)*** Models explaining the breaks in the intergenic colinearity after excision of DNA-Transposon.
STOW: DNA-Transposon (here *Stowaway* element), DSB: Double Strand Break, SSA: Single Strand Annealing, SDSA: Synthesis dependent Strand Annealing. For simplicity, the overhangs of the excision are not drawn. Dashed lines indicate deleted, gray lines inserted sequences, respectively. (**A**) The three different repair pathways after DNA-Transposon excision. The DSB can be repaired with a simple ligation, resulting in a classical footprint without change in the sequence. DSB repair by SSA is shown in (**B**) and repair by SDSA in (**C**). (**B**) DSB repair by SSA leads to breaks in intergenic colinearity through deletion. After species divergence, *TE_1* excises on *B. distachyon* and *TE_2* excises on *B. sylvaticum*. DSB repair via the SSA pathway leads to overlapping deletions, which after repair are seen as a break in intergenic colinearity. (**C**) DSB repair by SDSA leads to breaks in intergenic colinearity through insertion of filler DNA. *TE_1* excises in *B. sylvaticum*. SDSA repair resects the excision site by exonuclease activity. The break is repaired through filler DNA from another chromosome. This leads to a break in intergenic colinearity through insertion of non-homologous filler DNA.


A
Pathways of DSB repairs after the excision of DNA-Transposons
STOW
TA
AT
TA
AT
Simple ligation
SSA
SDSA
TA TA
AT AT
classic footprint
Deletion
Insertion of filler DNA
e.g. translocation
B
Model leading to the interuption of intergenic colinearity
by SSA after excision of DNA-Transposons
Ψ
TE_1
TE_2
B. distachyon
B. sylvaticum
Orthologous locus
Differential excision
of TEs
DSB repair by SSA
leads to deletions
deletion
Repair results in break
of colinearity
C
Model leading to the interuption of intergenic
colinearity by a combiantion of SDSA and SSA
after excision of one DNA-Transposon
Excision of TE in
B. sylvaticum
Exonuclease digestion at the
excision site leads to deletions
Repair by SDSA introduces
filler DNA at the excision site
colinear
filler DNA,
not colinear
colinear

first removes DNA past the boundary of the element and the DSB is subsequently repaired filler DNA via SDSA (Figure 3.5c). In this process, large fragments can be deleted on both sides of the DNA-Transposon. Alignments of these regions show a breakdown of sequence homology at the borders of the introduced filler DNA in one sequence, while the colinearity break in the other sequence will not be precisely at the borders of the element due to removal of flanking sequence through exonuclease activity (Figure 3.5c).

We discovered 51 cases where the excision of the transposon caused deletions on both sides. By simulating random distributions of InDels, we demonstrated that it is indeed highly likely that such InDels are the result of DNA-Transposon excision. Interestingly, similar DSB repair patterns of deletions an insertions after excision of *Mariner* elements have been reported in *C. elegans*, *Drosophila* and mouse (*Mus Musculus*) (Bryan *et al.*, 1990; Fischer *et al.*, 2001; Robert and Bessereau, 2007), indicating that excision and repair mechanisms are the same across kingdoms, at least for *Mariner* elements. However, these previous studies reported the deletions and/or insertions of at most a few base pairs (bps) at the excision site. In contrast, in our study, we observed deletion or insertion of dozens or hundreds of base pairs (bps), indicating that DSB repair in plants may be less precise than in animals.

DSB Repair Is an Important Driving Force of Genomic Turnover

"Genomic turnover" has been described as a balance between creation of new DNA through TE amplification and removal of DNA through unequal crossing-over and random deletions (Wicker *et al.*, 2003a; Vitte and Panaud, 2005). Many of the apparently random deletions have been described to result from what is generally referred to as "illegitimate recombination" (Devos *et al.*, 2002; Wicker *et al.*, 2003a). Excision of TEs followed by DSB repair through SSA is a good explanation for the observed illegitimate recombination signatures.

In all five loci, we aligned over 48 kilo base pairs (kb) of sequence between *B. distachyon* and *B. sylvaticum*. Excision of 111 DNA-Transposons (60 where the InDel precisely bordered the excision site and 51 with deletions on both sides) resulted in the removal of 23,531 bps (4.8% of the aligned sequences). Additionally, insertion of 5,089 bps filler DNA (approximately 1% of the sequence that can be aligned) via SDSA further contributed to the divergence of the sequences. Since the relatively recent divergence of the two species, approximately 5.9% of all sequences that can be aligned have been affected. Therefore, we conclude that DSB repair following transposon excisions is a major mechanism driving the reduction of sequence colinearity in the *Brachypodium* genomes, and possibly in plant genomes in general. Here one should note that a large portion of alignable sequences (310 kb) were coding regions of genes. These are obviously under strong selection pressure, as we found only one TE insertion disrupting an exon. Almost all TE activity therefore took place in intergenic or intronic regions. If one excludes the 310 kb of coding sequence (CDS), the TE excisions described here were responsible for the turnover of almost 17% of the sequences.

It must be emphasized that we do not claim that all random deletions are due to excision of DNA-Transposons. Indeed, we found deletions inside the 22 orthologous TEs that were not caused by TE excision and removed nearly 50% of the TE sequences. Although the sample size is small, these data indicate the existence of other important causes for deletions.

Indirect Influence of TEs on Genomic Colinearity

In a previous study, we showed that repair of DSBs caused by TE insertions and template slippage events can explain gene movement that leads to erosion of gene colinearity between distantly related species such as rice and *Brachypodium* (Wicker *et al.*, 2010). In the present study, we extend that model by providing explanations

for the short-term erosion of intergenic colinearity between closely related species. We propose that the main impact of TEs on genomic colinearity is indirect, because the change is caused by the repair of DSBs that results from their activity, mainly excision. This is consistent with a Southern blot analysis performed in maize (*Zea mays*), showing that *Ac/Ds* excision at the *P* locus induced homologous recombination (Athma and Peterson, 1991) and insertion of *Mutator* elements at the *KNOTTED-1* locus leads to increased intrachromosomal recombination (Lowe *et al.*, 1992).

Here, we described how repair of DNA-Transposon-induced DSBs corrodes colinearity in the relatively small genome of *B. distachyon*. Comparisons within larger genomes such as sorghum, wheat, barley or maize are required to further investigate the mechanisms of genomic turnover. In particular, the impact of other TE families in highly repetitive genomes needs to be addressed, as these genomes contain TE families with thousands of copies while Miniature Inverted Repeat Transposable Elements (MITEs) by far outnumber all other TE types in *Brachypodium*. Studying their role of these TE families in genomic rearrangements will be essential for our understanding of genome evolution in different species.

3–4 Experimental Procedures

Sequence Analysis

Screening of the *B. sylvaticum* bacterial artificial chromosome (BAC) library (6.6 coverage, Foote *et al.* 2004) was performed as previously described by Bossolini *et al.* (2007). One overlapping BAC was identified by chromosome walking using bordering genes as probes for screening. Sanger sequencing of BACs and `PhredPhrap` sequence assembly was done as previously described (Bossolini *et al.*, 2007; Wicker *et al.*, 2007b). Remaining gaps between sequence contigs were closed with PCR to confirm their correct linear order.

For sequence analysis, `BLAST` (Altschul *et al.*, 1997), `EMBOSS` (Rice *et al.*, 2000) and `DOTTER` (Sonnhammer and Durbin, 1995) were used. *B. distachyon* genomic and protein coding sequences libraries were obtained from `http://files.brachypodium.org`. Newly identified genes were found by `BLASTN` searches against rice and *B. distachyon* CDS libraries. To identify coding capacities in the conserved non-coding sequences we performed additional `BLASTX` searches against known genes, transposable elements (TEs), tRNA and rRNA libraries from plants. For annotation and reannotation of *B. distachyon* and *B. sylvaticum* genes, *B. distachyon* and rice CDS with the highest identity were aligned with the *B. distachyon* or *B. sylvaticum* genomic sequence using `DOTTER` to determine the positions of introns, exons, start and stop codons. Repetitive elements were annotated with `DOTTER` using a repeat database which is available upon request and classified as proposed in Wicker *et al.* (2007a).

For identification of previously non-identified *Stowaway* elements, a `Perl` program was written which extracted putative elements, allowing one mismatch in the recognition motif. tRNA genes were identified using Arabidopsis tRNA databases for `BLASTN` searches. *De-novo* gene prediction and microRNA prediction

were performed using RiceGAAs (ricegaas.dna.affrc.go.jp) and RNAspace (rnaspace.org), respectively, after masking all annotated genes and TEs. *De novo* gene prediction did not identify new genes in addition to those that were already identified based on homology.

To avoid the alignment of long and non-homologous sequences, we created smaller alignments that were subsequently reconstructed. To determine the borders of homologous and non-homologous regions, we manually analyzed dot–plot alignments of two orthologous loci to determine large breaks in colinearity. Those regions were aligned independently and reassembled to produce a sequence alignment for the entire loci.

Divergence Time Estimates

All divergence time estimates were calculated using a mutation rate of 1.3×10^{-8} substitutions per site per year. A Perl program was developed which automated the procedure of extracting intergenic sequences of colinear regions. These were then aligned with the program water from the EMBOSS package and the nucleotide substitutions in the aligned sequences were counted. The Kimura 2-parameter criterion was applied to determine the transition to transversion ratio. Alignments positions which contained a transition from C to T in CG and CNG sites were removed to exclude a bias which could be introduced by DNA methylation. The age of LTR retrotransposons was estimated as described by SanMiguel *et al.* (1998). For divergence time estimates using CDS of genes, we used only alignment positions corresponding to the third codon base of codons for Ala, Gly, Leu, Pro, Arg, Ser, Thr and Val. For Leu, Arg and Ser (which all have 6 possible codons), we used only the codons starting with CT, TC and CG, respectively. These are the codons in which the third base can be exchanged without causing an amino acid change. As the third codon base is redundant, it offers a very good way to exclude evolutionary pressure

since its exchange does not alter the amino acid. All `Perl` scripts used in this study are available upon request.

Development of Evolutionary Models and Simulation of Distribution of Deletions

For development of evolutionary models (such as DSB repair events following TE excisions) we applied the following two rules. First, it is extremely unlikely that two independent events (e.g a TE insertion, or a deletion) take place at the exact same base pair position. Second, the explanation that uses the smallest number of evolutionary steps is preferred over explanations that include more steps. This leads to selection of a likely model, but we do not exclude the possibility that other, albeit less likely, scenarios could lead to the same result.

To test whether the observed distribution of deletions is associated with TE activity, a `Perl` program was written to simulate random distributions of deletions. To match the natural situation as closely as possible, the number and sizes of deletions randomly distributed in a locus corresponded to the observed data for the respective locus. In other words, we randomly distributed the same number of deletions with the same size distribution in the intergenic regions of a locus (e.g. if one species locus contained three deletions of 100, 200 and 300 bp, respectively, we introduced the same set of deletions randomly in the other species). To distribute the deletions on the respective locus we used the following restrictions: (i) the simulated deletion has to lie within the regions that could be aligned between *B. sylvaticum* and *B. distachyon*, (ii) the simulated deletion is not allowed to lie inside an exon of a gene (to reflect that only very few deletions were found in exons) and (iii) overlapping simulated deletions are not allowed (i.e. a deletion can not be placed in a region where another one was already placed). These restrictions guaranteed that the same number of deletions with the same size distribution would be placed

on the respective locus. The procedure was repeated 1000 times for each locus. For each repetition, we determined, how many deletions removed entire TEs, how many only partially overlapped TEs, and how many are outside TEs. Compilation of the results from the 1000 repetition gave us the values expected from a random distribution. We determined if the observed data was significantly different from the expected with a χ^2 test.

Sequence Deposition

The complete sequence data of the *All* locus this article has been deposited in the GeneBank database under following accession numbers: HE650836. The dataset can also be obtained from the authors upon request.

3–5 Acknowledgments

This study was supported by the Swiss National Science Foundation Grant 31003A-122242. JPB carried out lab work, data analysis and wrote the paper. TM and NS funded and carried out sequencing of the *Sdw3* locus. BK funded and supervised sequencing of the *All* locus and reviewed and edited the manuscript. TW conceived and coordinated the study and reviewed and edited the manuscript.

Methodological aspects of divergence time estimates*

4

Abstract

The estimation of divergence times is used to determine when two species diverged from their common ancestor. This requires the knowledge of the molecular phylogeny based on genomic or protein sequences and a fossil reference to calibrate the substitution rate. Depending on the type of sequence used for divergence time estimation, e.g. genic or intergenic, the appropriate substitution rate has to be used. Here. we describe an approach to estimate divergence times using only one substitution rate, independently of the analyzed sequence. The basic idea was to use only sites which are neutral, i.e. masking non-synonymous sites in CDS sequences and removing putative C→T methylation sites in intergenic sequences. These methods were tested between the closely related species *Brachypodium distachyon* and *Brachypodium sylvaticum*, which diverged between 1.7–2.0 million years ago (MYA). The results, using assumed neutral sites in CDS and intergenic sequences, showed a high similarity. In addition, we present a method to determine a minimal divergence time using non-orthologous LTR-retrotransposons where we removed putative methylated sites, resulting in an estimated minimal divergence time of

*Part of the presented data has been published in Buchmann *et al.* (2012)

approximately 2.4 MYA.

4–1 Introduction

Divergence time estimates are used to estimate the time when two species diverged from a common ancestor. After divergence from a common ancestor, the sequences in the new species accumulate mutations independently from each other. Aligning sequences, e.g. genic or intergenic, from orthologous loci allows to quantify the change in nucleotide composition. This difference, sometimes called distance, between two nucleotide sequences combined with a substitution rate can be used to estimate the divergence time of two loci. The divergence time between two loci is linked to the divergence time of the two species from where they were extracted, allowing the estimation of divergence times without knowing the complete genomic sequence.

Generally, two approaches are used to estimate the divergence time of species: the molecular clock of genes and the analysis of synonymous sites. Using the molecular clock assumes that accumulation of changes on the DNA occurs at a roughly constant rate over time. Therefore, two species after divergence amass substitutions at a similar rate. This rate differs between taxonomic groups: e.g., higher primates and some birds have a slower rate than rodents or Drosophila (Britten, 1986). Because the substitution rate also differs for individual genes or gene families, the knowledge of specific rates for each gene is required.

The analysis of synonymous sites uses only sites which are presumably free from evolutionary pressure, i.e. sites which, when changed, encode the same amino acid. This is mostly the third base pair of codons in coding sequence (CDS) sequences. To estimate divergence times using synonymous sites, a basic substitution rate is required which is the same for all synonymous sites.

Counting the substitutions between genes used for the molecular phylogeny

and knowing the divergence time of the species from which the genes were analyzed allows to calculate a substitution rate. Once the substitution rate is known, divergence times of other species can be estimated by comparing DNA or protein sequences. The calibration of a substitution rate needs i) a robust molecular phylogenetic tree and, ii) accurate fossil dating for at least one node in this tree. Molecular phylogenetic data can be obtained through phylogenetic analysis of genes which are ubiquitous among species, e.g. the genes ribulose-bisphosphate carboxylase (*rbcL*, Duvall *et al.* 1993) or alcohol dehydrogenase (*Adh*, Goloubinoff *et al.* 1993; Gaut *et al.* 1996) in plants. Paleontological data is used to estimate the divergence time between two species, whereby the minimal time of divergence can be estimated by the earliest fossil finds (Marshall, 1990; Goloubinoff *et al.*, 1993; Ayala *et al.*, 1998). The determination of the maximal divergence time is more complicated since one has to distinguish the ancestor from a possible sister line (Marshall, 1990). In grasses (monocotyledons), fossil pollen has been the primary fossil source (Daghlian, 1981). However, one has to keep in mind that divergence times derived from paleontological samples as well as the phylogenetic molecular data can be erroneous or tainted with a broad standard error (Sanderson and Doyle, 2001; Graur and Martin, 2004).

The *Adh* loci in grasses harbor genes from the multigene family encoding alcohol dehydrogenases and have been extensively analyzed to calculate substitution rates. Analysis of those loci in palm and grasses resulted in the first calculation of a synonymous substitution rate in grasses, 6.5×10^{-9} substitutions per site per year, (Gaut *et al.*, 1996). Ma and Bennetzen (2004) analyzed approximately 1 mega base pair (Mbp) of orthologous loci between the two rice species *Oryza sativa* and *Oryza glaberrima.* Applying the substitution rate of 6.5×10^{-9} substitutions per site per year on synonymous sites from 24 genes they estimated the divergence between *Oryza sativa* and *Oryza glaberrima* to be approximately 0.44 MYA. However, they noted that the nucleotide substitution in intergenic regions is approximately 2-fold

higher than in genes at the *Adh* loci. In addition, the insertion times for a number of non-orthologous LTR-retrotransposons using 6.5×10^{-9} substitutions per site per year was older than the divergence time of the two species. Using the proposed substitution rate derived from the intergenic sequences to date the insertion times placed all non-orthologous LTR-retrotransposons after species divergence. Therefore, Ma and Bennetzen (2004) proposed a higher rate for synonymous substitutions of 1.3×10^{-8} substitutions per site per year . This rate has since been used most frequently in plants and fungi.

The differences in the two types of sequences, genic and intergenic/TE, has consequences when used for divergence time estimates. Transposable elements TEs in plants are more likely to be methylated than genic regions (SanMiguel *et al.*, 1998; Bennetzen *et al.*, 1994). In methylated DNA, the C nucleotide at 5′-CG-3′ or 5′-CNG-3′ sites can convert spontaneously into a T, resulting in 5′-TG-3′ and 5′-TNG-3′. Therefore, in intergenic or transposable element (TE) sequences, the possibly methylated sites should be excluded because they have an accelerated mutation rate. In CDS sequences, only synonymous sites in codons should be used because the exchange of a synonymous sites does not change the encoded amino acid. These sites are presumably free from selection pressure and accumulate mutations at the basic rate of 1.3×10^{-8} substitutions per site per year . Therefore, removing the sites which are likely under evolutionary pressure from all analyzed sequences, we can use the substitution rate proposed by Ma and Bennetzen (2004) (1.3×10^{-8} substitutions per site per year).

Here we present procedures which were used to estimate divergence times for genic and intergenic/TE sequences by examining only synonymous sites in protein coding sequences and removing potentially methylated sites in intergenic sequences. We used this approach in Buchmann *et al.* (2012) to date the divergence between *Brachypodium distachyon* and *Brachypodium sylvaticum*. The five loci

used for this analysis were described in Buchmann *et al.* (2012) (Chapter 3) and the following analysis is a more detailed description of the divergence time estimates used in that study.

4–2 Results

To estimate the divergence time between *B. distachyon* and *B. sylvaticum* we used CDS sequences from colinear genes as well as the intergenic sequences between those genes. We used a dataset consisting of five orthologous loci which are described in more detail in Chapter 3 and Buchmann *et al.* (2012). Both types of sequences were aligned using the same parameters. We removed the potentially methylated sites in intergenic sequences while in CDS sequences we used only nucleotides which were presumed neutral, i.e. the nucleotides in the codon sequence which are unaffected by selective pressure. We compared the impact on divergence time estimates using two datasets for each type of sequence. In one dataset, we used all sites while in the second only non-methylated or synonymous sites were analyzed. Analysis of non-synonymous sites has been included for the sake of completeness.

The Quality of the Sequence Alignment Is Crucial for Analysis

The orthologous intergenic sequences from *B. distachyon* and *B. sylvaticum* were aligned using the program `water` from the `EMBOSS` package (Rice *et al.*, 2000). To reduce the influence of sequences possibly under selection (up- and downstream regulatory sequences) we removed 1 kilo base pair (kb) from the 5′ and 3′ ends of the sequence.

We aimed for sequence alignments containing few but long gaps. For example, transposable elements (TEs) usually produce clear borders and do not unravel in sequences with low similarity, therefore resulting in rather long blocks of aligned sequences which are not interrupted by small blocks of not aligned sequences. Such rigorous alignments are based on very strict parameters for the sequence alignments: large gap opening penalties but low gap extension penalties. We choose a gap opening penalty of 30 and a gap extensions penalty of 0.1 (default: gap open penalty

= 10, gap extension penalty = 0.5). Nevertheless, some regions were still poorly aligned, showing dispersed and small blocks of aligned sequences (approximately 10–20 base pair (bp) long) with low similarity. Therefore, all alignments were checked manually for poorly aligned regions which were then removed. For the analysis, the aligned segments from each locus were concatenated, creating one long alignment for each locus and reducing the standard error for the subsequent divergence time estimates.

Removal of Potentially Methylated Sites in Intergenic Regions Reduces the Estimated Divergence Times Approximately 12%

In methylated intergenic sequences, spontaneous methylation of 5′-CG-3′ and 5′-CNG-3′ sites can occur which leads to over-estimated divergence times. To overcome this problem, we wrote the `Perl` program `rmMethylPos.pl` which screened the sequence alignments for `CG` and `CNG` sites. If such a site is present, the program checks if the aligned nucleotide in the other sequence is `T`. If yes, we removed this position from the alignment. The unaltered dataset was named C^m while the dataset where we removed the putative methylated sites was named C^{nm}.

From the five loci we aligned between 3.3 and 30.8 kb of intergenic sequences on which we identified 16 orthologous DNA-Transposons but no orthologous retrotransposons. The estimated divergence times (EDTs) are between 2.856±0.081 and 3.742±0.219 million years (Myr) for the C^m sequences while C^{nm} sequences showed younger EDTs that were between 2.519±0.076 and 3.426±0.207 Myr (Table 4.1, Figure 4.1). In average, the EDTs between the two datasets differed 0.35 Myr, approximately 12%. We conclude that removal of potentially methylated sites has a notable influence on EDTs. Therefore, we used only the C^{nm} dataset for all further EDTs calculations from intergenic sequences.

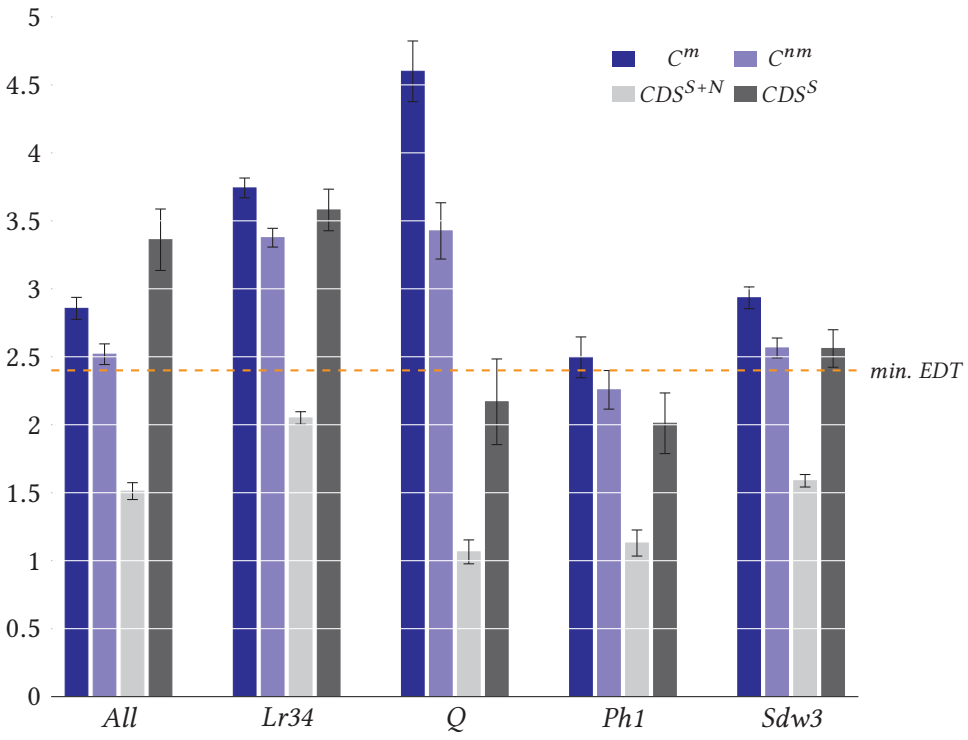


Figure 4.1 Bar graph for estimated divergence times using intergenic and CDS sequences.
The loci are indicated at the x-axis with EDTs plotted in following order: C^m, C^{nm}, CDS^{S+N}, CDS^S, as indicated in the legend. The error bars indicate the standard error. The *min. EDT* indicates the minimal divergence time (2.4 MYA) which was derived from LTR insertion times.

Table 4.1 Differences in divergence time estimates of *B. distachyon* and *B. sylvaticum* using intergenic sequences in five genomic regions. C^m: methylated dataset; C^{nm}: non-methylated dataset; Δ: differences between methylated and non-methylated EDT; loc: locus as described in Buchmann *et al.* (2012); si: aligned bp for whole locus; ti: transitions; tv: transversions; EDT: Estimated divergence time ± standard error; Myr: million years.

		C^m				C^{nm}				Δ		
loc	tv	si [bp]	ti	ti:tv	EDT [Myr]	si [bp]	ti	ti:tv	EDT [Myr]	ti	ti:tv	EDT [Myr]
All	642	18,228	649	1.01	2.856 ±0.081	18,069	490	0.76	2.519 ± 0.076	159	0.25	0.337
Lr34	1,483	30,871	1,330	0.90	3.742 ±0.073	30,594	1,050	0.71	3.376 ± 0.069	280	0.19	0.366
Q	162	3,337	142	0.88	3.742 ±0.219	3,311	116	0.72	3.426 ± 0.207	26	0.16	0.316
Ph1	131	4,546	151	1.15	2.496 ±0.150	4,519	124	0.95	2.257 ± 0.142	27	0.20	0.239
Sdw3	684	19,358	718	1.05	2.934 ±0.080	19,180	539	0.79	2.565 ± 0.073	179	0.26	0.369
Total	3,102	76,340	2,990	0.96	3.250 ±0.042	75,673	2,319	0.75	2.896 ± 0.038	671	0.22	0.354

EDTs Derived from Modified Intergenic and CDS Datasets Are Almost Identical

Protein coding sequences (CDS) are under a higher selection pressure than intergenic sequences. Using those sequences without any modification would, in contrast to methylated intergenic sequences, lead to underestimated divergence times. One way to reduce the effect of selection pressure is to use only synonymous sites because they are supposed to be neutral. The codons where the third nucleotide can be exchanged without changing the encoded amino acid are `Ala`, `Gly`, `Val`, `Pro`, `Thr`. The amino acids `Arg`, `Ser`, `Leu` have each 6 possible codons. whereby we used only codons which started with `CT` (`Leu`), `TC` (`Ser`) and `CG` (`Arg`). A `Perl` program was developed which screens the CDS alignments and masked the codons which did not correspond to those described above. This dataset was designated CDS^{S} and the unaltered dataset was designated CDS^{S+N}. We analyzed in total 105,761 bp of CDS sequences, ranging from 5,266 to 32,127 bp. As expected, the EDTs of the CDS^{S+N} set are much more recent, being approximately two times lower than the EDTs from the CDS^{S} set (Table 4.2) and were not further analyzed.

We focused on the comparison between the two datasets C^{nm} and CDS^{S} because they are based on sequences containing only sites which are assumed to be neutral. In fact, the comparison of the EDTs derived from the two datasets show very similar values. In three (*Lr34*, *Ph1*, *Sdw3*) out of the five loci we estimated divergence times which are within each others standard error whereby in the case of *Sdw3* the EDT is almost identical. This indicates that in those cases no strong difference in the EDTs was found. This is supported by comparing the average EDT from the two altered datasets which show very similar divergence times.

It is expected that the estimates derived from the CDS^{S} dataset tend to be more recent than those from the C^{nm} dataset. However, we observed the opposite in

Table 4.2 Differences in divergence time estimates of *B. distachyon* and *B. sylvaticum* using CDS sequences in five genomic regions. CDS^{S+N}: all sites; CDS^{S}: only d_S sites; Δ: differences between all and d_S sites; loc: locus as described in Buchmann *et al.* (2012); si: aligned bp for whole locus; ti: transitions; tv: transversions; sub: substitutions; EDT: Estimated divergence time ±standard error; Myr: million years.

	CDS^{S+N}					CDS^{S}					Δ		
loc	si [bp]	ti	tv	ti:tv	EDT [Myr]	si [bp]	ti	tv	ti:tv	EDT [Myr]	si [bp]	si [%]	EDT [Myr]
All	16,875	346	229	1.51	1.512 ±0.062	2,768	104	124	0.84	3.361 ± 0.226	14,107	16.4	1.849
Lr34	41,641	1,167	972	1.20	2.050 ±0.046	6,613	292	286	1.02	3.580 ± 0.153	35,028	15.8	1.530
Q	5,266	81	62	1.31	1.065 ±0.088	885	25	23	1.09	2.169 ± 0.315	4,381	16.8	1.104
Ph1	9,852	142	142	1.00	1.130 ±0.096	1,624	31	51	0.61	2.011 ± 0.223	8,228	16.5	0.881
Sdw3	32,127	678	612	1.11	1.588 ±0.046	5,562	166	188	0.88	2.561 ± 0.138	26,565	17.3	0.973
Total	105,761	2,414	2,017	1.20	1.661 ±0.027	17,452	618	672	0.92	2.996 ± 0.084	88,309	16.5	1.335

the *All* locus as the EDT is more recent in the C^{nm} than in the CDS^{S} dataset. This suggests that either intergenic sequences in the *All* locus change in a slower pace than CDS^{S} sequences or other mechanisms keep the intergenic sequences more conserved. The *Q* locus showed by far the most recent divergence time and we assume that this discrepancy could be explained by the small alignment length. Nevertheless, we derived very similar EDTs from the C^{nm} and CDS^{S} datasets. This supports our approach of using only sites in sequence alignments which are assumed to be neutral.

LTR-retrotransposons in *B. sylvaticum* and *B. distachyon* inserted 0.0–3.67 MYA

After insertion in a new location, the identical LTRs of a retroelement start to accumulate mutations independently from each other. The accumulated mutations in the two LTRs of one element can be analyzed through sequence alignments, thereby counting substitutions which then can be used to determine the time of insertion. We identified 21 non-orthologous full length LTR elements in four of the five analyzed loci, i.e. they inserted after species divergence (Buchmann *et al.*, 2012). Complete LTR elements (which allow this type of molecular dating) were identified on both sequences at the loci *Lr34* and *Sdw3*. In contrast, at the loci *Ph1* and *All* complete LTRs were identified only on *B. distachyon* sequences while no complete LTRs have been identified at the *Q* locus.

Using dot–plot alignments, we defined the borders of the LTRs from the 21 non-orthologous LTR elements. We aligned the two LTRs from each elements and counted the transitions and transversion. Similar to the intergenic sequences, two datasets were produced. One dataset was not altered (LTR^{m}) while in the second dataset the methylated CG and CNG sites were removed (LTR^{nm}). The insertion time was estimated for each element on each sequence and locus, whereby for the latter

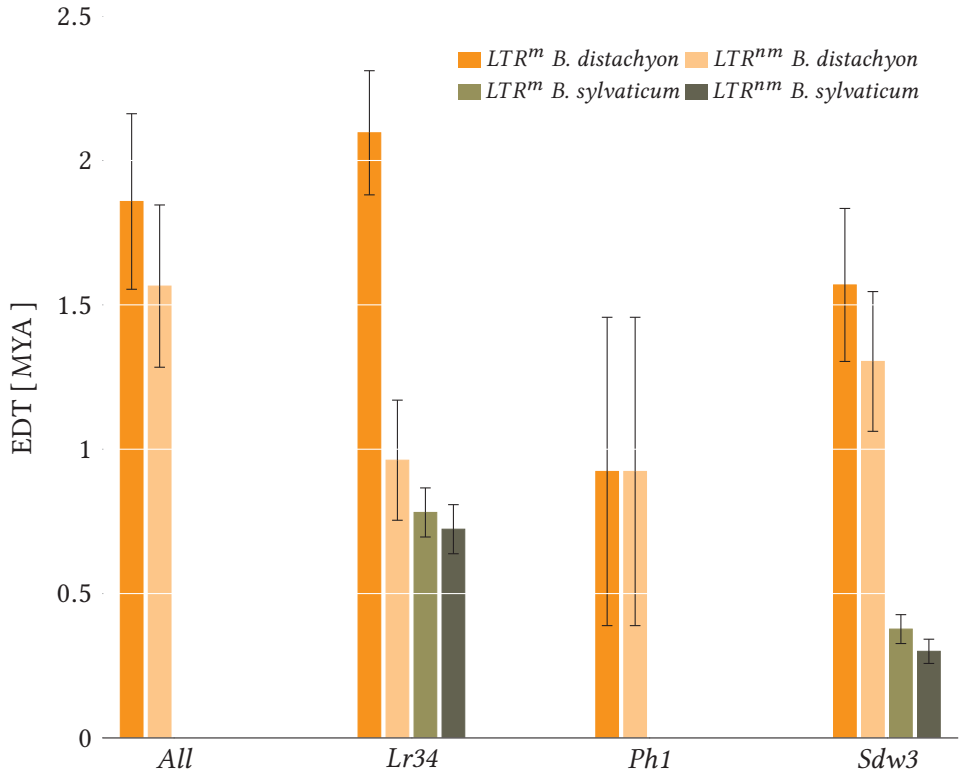


Figure 4.2 Bar graph for LTR insertion time estimates in the loci *All*, *Lr34*, *Ph1* and *Sdw3* in *B. sylvaticum* and *B. distachyon*.
The bars are colored as described in the legend whereby the first two bars indicate the insertion times which were estimated in the *B. sylvaticum* sequences from the LTR^m and LTR^{nm} dataset, respectively, while the last two bars indicate the same datasets in *B. distachyon*. The error bars indicate the standard error. No complete LTRs have been found in the *B. sylvaticum* sequences at the loci *All* and *Ph1*.

the individual transitions and transversions were summed up (Table 4.3, Figure 4.2).

In the LTR^m dataset insertion times between 0±0.0 and 3.142±0.396 MYA were estimated while in the LTR^{nm} dataset between 0±0.0 and 2.473±0.350 MYA. As mentioned above, the results from the LTR^{nm} dataset were considered more reliable. The *Lr34* locus on the *B. sylvaticum* sequence harbors the most recent insertions while on the *B. distachyon* sequence we identified the oldest insertions at the *Sdw3* locus.

LTR-retrotransposon Insertions in *B. sylvaticum* are younger than *B. distachyon*.

Table 4.3 Insertion times of non-orthologous LTR retrotransposons in *B. distachyon* and *B. sylvaticum* to estimate a lower limit for the divergence time of the two species. Entries marked with * denote insertion times which standard error does not overlap. EDT: Estimated divergence time ± standard error; si: analyzed sites; ti: transitions; tv: transversions; LTR^m: methylated LTR dataset; LTR^{nm}: LTR dataset with removed methylated sites.

				LTR^m				LTR^{nm}				Δ	
Locus	Species	LTR	tv	si	ti	ti:tv	EDT[MYA]	si	ti	ti:tv	EDT[MYA]	ti	EDT[MYA]
Ph1	***B. distachyon***	*RLC_1*	2	127	1	0.50	0.923 ±0.534	127	1	0.50	0.923 ±0.534	0	0.000
All	***B. distachyon***	*RLG_C158*	9	403	7	0.78	1.569 ±0.396	401	5	0.56	1.377 ±0.396	2	0.192
		RLC_C193	6	411	16	2.67	2.150 ±0.465	407	12	2.00	1.761 ±0.419	4	0.389
		Total	15	814	23	1.53	1.858 ±0.304	808	17	1.13	1.565 ±0.281	6	0.293
Lr34	***B. distachyon***	*RLG_2**	29	842	36	1.24	3.142 ±0.396	828	22	0.76	2.473 ±0.350	14	0.669
		RLC_10	5	340	7	1.40	1.392 ±0.403	339	6	1.20	1.277 ±0.388	1	0.115
		RLC_3	4	707	18	4.50	1.230 ±0.265	703	14	3.5	0.996 ±0.238	4	0.234
		Total	38	1,889	61	1.60	2.096 ±0.215	1,870	42	1.11	1.962 ±0.208	19	0.134
	B. sylvaticum	*RLG_1*	4	884	21	5.25	1.115 ±0.226	882	20	5.00	1.069 ±0.219	1	0.046
		RLG_2	9	774	5	0.56	0.703 ±0.188	772	3	0.33	0.603 ±0.173	2	0.100
		RLG_3	0	773	2	–	0.100 ±0.069	773	2	–	0.100 ±0.069	0	0.000
		*RLC_Mara_1**	13	1,073	22	1.69	1.284 ±0.219	1,070	19	1.46	0.176 ±0.207	3	1.108
		RLC_4	2	462	5	2.50	0.588 ±0.223	462	5	2.50	0.588 ±0.223	0	0.000
		RLC_C340	0	183	0	–	0.000 ±0.000	183	0	–	0.000 ±0.000	0	0.000
		Total	28	4,149	55	1.96	0.781 ±0.085	4,142	49	1.75	0.723 ±0.085	6	0.058
Sdw3	***B. distachyon***	*RLG_2**	4	468	5	1.25	0.500 ±0.203	465	2	0.50	0.250 ±0.142	3	0.250
		RLC_2	12	201	6	0.50	3.669 ±0.880	199	4	0.33	3.288 ±0.834	2	0.381
		RLC_1	4	213	4	1.00	1.484 ±0.526	212	3	0.75	1.115 ±0.457	1	0.369
		Total	20	882	15	0.75	1.569 ±0.265	876	9	0.45	1.304 ±0.242	6	0.265
	B. sylvaticum	*RLC1_1**	3	769	4	1.33	0.353 ±0.134	766	1	0.33	0.200 ±0.100	3	0.153
		RLC3_1	3	427	11	3.67	1.296 ±0.350	425	9	3.00	1.111 ±0.323	2	0.185
		RLC3_2	1	440	4	4.00	0.442 ±0.200	439	3	3.00	0.353 ±0.176	1	0.089
		RLG1_1	1	3,513	11	11.00	0.130 ±0.038	3,510	8	9.00	0.100 ±0.346	3	0.030
		RLG2_1	1	750	17	17.00	0.946 ±0.226	747	14	14.00	0.788 ±0.203	3	0.158
		RLG3_1	1	282	3	3.00	0.550 ±0.276	282	3	0.33	0.550 ±0.276	0	0.000
		Total	10	6,181	50	5.00	0.377 ±0.050	6,169	38	3.80	0.300 ±0.042	12	0.077

The insertion times of LTR-retrotransposons in the *B. sylvaticum* sequences were found to be more recent than those on the *B. distachyon* sequence (Table 4.3). The estimated insertion times on the *B. distachyon* and *B. sylvaticum* sequences at the locus *Sdw3* in the LTR^{nm} dataset differ considerably. Only one of six estimated insertion times is older than 1 MYA in the *B. sylvaticum* sequences at the *Sdw3* locus. In contrast, the *B. distachyon* sequence at this locus harbors only one of three estimated insertion which is younger than 1 Myr. The same can be observed for the *Lr34* locus.

LTR-retrotransposon Insertion Times Set Lower Limit for the EDT to 2.4 MYA

Analyzing the insertion time of non-orthologous LTR elements allows to estimate a lower limit for the estimated divergence time because those elements have transposed after the divergence of *B. sylvaticum* and *B. distachyon*. Due to the lifestyle of Class I elements a copy is retained at the original location. In case of non-orthologous elements, their insertion time cannot be older than the divergence time of the host genomes, in our case *B. sylvaticum* and *B. distachyon*. The insertion times from all 21 LTR elements are between 0–3.288 Myr (Table 4.3). The youngest insertion (0±0.0 MYA) is the element *RLC_C340* in *B. sylvaticum* at the *Lr34* locus. The oldest inserted element, dating 3.288±0.834 Myr, was *RLC_2* in *B. distachyon* at the *Sdw3* locus. The majority, 18 out of 21 elements, inserted less than 2 MYA. We identified 11 elements which have inserted 1 MYA while 7 elements have inserted between 1–2 MYA. Only three elements have insertion times between 2–4 MYA.

The minimal insertion time is assumed to be the oldest inserted element, which is *RLC_2* in *B. distachyon* at the locus *Sdw3*. However, it has the largest standard error and is the only element which inserted between 3–4 MYA. The second oldest insertion has been identified for *RLG_2* (2.473±0.350 MYA) on the *B. distachyon*

sequence of the *Lr34* locus. It has a standard error which is two fold lower than for *RLC_2*, i.e. the oldest identified insertion. Due to the large standard error of *RLC_2* and its estimated insertion time, which is far off compared to the other estimates, we treated it as outlier. Most EDTs derived from the intergenic as well as CDS sequences were between 2–3 MYA. The second oldest insertion time of 2.473±0.350 MYA from *RLG_2* fits in this time range. We therefore propose that *RLG_2* represents the oldest estimated insertion time. Choosing the insertion time of *RLG_2* as the minimal estimated divergence time is supported by the youngest EDT for the *Ph1* locus, derived from the CDS^S dataset. The estimated divergence time for the *Ph1* locus was 2.011±0.223 MYA. Taking into account the standard error from *RLG_2* of 0.350 MYA, the EDT of the *Ph1* locus lays within the insertion time of *RLG_2*, linking the minimal insertion time with the most recent EDT.

4–3 Discussion

We used the dataset of five orthologous loci in *Brachypodium distachyon* and *Brachypodium sylvaticum* (described in Buchmann *et al.* (2012), Chapter 3) to analyze two different approaches for estimating divergence times. In the first approach we used intergenic while in the second one protein coding sequences (CDS). Both datasets were compared in two states: i) in a raw state, not considering the effects of methylation or non-synonymous base substitutions. Those datasets were designated C^m intergenic and CDS^{S+N} for the CDS sequences, respectively. ii) in a processed state where those effects were taken into account and corresponding sites were removed or masked for divergence time estimates. The datasets were designated C^{nm} for intergenic and CDS^S for protein coding sequences. The coding sequence (CDS) and intergenic datasets were then evaluated using the Kimura-2 parameter method to estimate the divergence time. In addition, the minimal divergence time for the two species was estimated using the insertion times from non-orthologous Long Terminal Repeat (LTR) elements. This dataset was analyzed in a similar way as the intergenic sequences which resulted in a a raw (LTR^m) and in a processed (LTR^{nm}) dataset where putative methylated sites were removed. The LTR dataset was used to determine a lower limit for the divergence time estimates.

Processed Datasets Result in Similar Divergence Time Estimates but Require Longer Sequence Alignments

The estimated divergence times (EDTs) derived from the C^{nm} and CDS^S dataset are very similar, despite the fact that one dataset is derived from intergenic while the other from coding sequences. This shows that using synonymous sites in CDS or masked putatively methylated sites in intergenic sequences allows to estimate divergence times using the same substitution rate. Thus, in any case the LTR^{nm} and CDS^S dataset should be used for dating. Our data indicate that at least 5,000 base

pair (bp) of aligned and processed sequence are required for robust divergence time estimations as the loci with the highest standard errors were *Q* and *Ph1* which have approximately 4–9-fold less sequence information compared to the other loci. In addition, they have the highest standard errors, indicating that these two loci have a high probability of being statistical outliers. In contrast, the CDS^S dataset for the locus *Sdw3* aligned 5,562 bp with the lowest standard error and the EDT is 2.5 million year ago (MYA), very close to the estimated minimal divergence time derived from non-orthologous retrotransposons. Only at the *All* locus the estimates for CDS and intergenic sequences differed strongly.

The differences of the EDTs between the five loci could be also a result of analyzing different haplotypes. In wheat and barley, different studies found sudden breaks at different loci which separated highly conserved from less conserved sequences (Isidore *et al.*, 2005; Scherrer *et al.*, 2005; Wicker *et al.*, 2009a). This suggests that parts of ancient haplotypes can still be detected today, which can influence the estimation of divergence times as some loci can be located on older haplotypes than others.

What causes the younger EDT in the C^m dataset compared to the CDS^S of the *All* locus is not clear. A possibility is the problem of so-called deep paralogs. The comparison of the *All* locus would then be the comparison of two paralogs, both present in the common ancestor but with different copies lost in *B. sylvaticum* and *B. distachyon*. To check for deep paralogs, additional orthologous *All* sequence from grasses could be compared.

Determination of LTR-retrotransposon Insertion Times Can Be Used to Estimate Minimal Divergence Times

We used the insertion time from non-orthologous LTR elements to estimate a lower limit for the divergence times. Using the LTR^{nm} dataset, we estimated that all 21

LTR elements were inserted between 0–3.6 MYA, of which two insertion estimates ranged from 3–4 MYA while most EDT derived from CDS^S and C^{nm} were between 2–3 MYA.

The minimal divergence time is expected to be close to the oldest insertion time which is 3.288 MYA. This estimate, however, showed two characteristics of a statistical outlier. First, the estimate is between 3–4 MYA, whereby most loci showed an EDT between 2–3 MYA. Second, it has the largest standard error in the dataset. Considering this estimate an outlier, the minimal estimated insertion time was derived from the element *RLG_2* on the *B. distachyon* sequence at the *Lr34* locus which inserted 2.4±0.350 MYA. This is in accordance with the estimated divergence times from intergenic and CDS sequences which were mostly between 2–3 MYA.

Recent Insertions of LTR-retrotransposons in *B. sylvaticum* Sequences

At the two loci *Lr34* and *Sdw3* we analyzed the insertion times of complete LTR elements in *B. sylvaticum* and *B. distachyon*. Interestingly, the estimated insertion times of LTR-retrotransposons between the two species showed a much wider difference. EDTs in the *B. sylvaticum* sequences are mostly below 1 MYA. In contrast, the EDT in *B. distachyon* are mostly above 1 MYA. In addition, we found twice the number of LTR elements on the *B. sylvaticum* sequences (12) than in *B. distachyon*. This suggests that the loci *Lr34* and *Sdw3* in *B. sylvaticum* had a higher LTR activity than in *B. distachyon* since species divergence. The genome size of *B. sylvaticum* is comparable to that of rice (Foote *et al.*, 2004), which is approximately 1.5 times larger than the of *B. distachyon*. The recent activity of LTR-retrotransposons could have been one reason which led to the size differences between the two genomes.

Compensating for Codon Usage Bias

To use only one substitution rate among several types of sequences we removed or masked the sites which were assumed to be under selective pressure (e.g. non-synonymous sites) or have changed spontaneously (e.g. methylated sites in intergenic sequences). However, we did not account for codon usage bias. Codon usage bias describes the selection on synonymous sites, e.g. the preference for a certain codon over another. This phenomena has been described in *Drosophila melanogaster* where the preference for certain codons increases the translational accuracy (Akashi, 1994). Codon usage bias has also been observed in mammals, where mRNA stability and splicing is affected by the selection of certain synonymous sites (Chamary *et al.*, 2006). In bacteria, Lafay *et al.* (1999) reported different codon usage bias for two strains while in *Heliobacter pylori* codon usage bias has not been detected at all (Lafay *et al.*, 2000). Codon usage bias has also been reported in plants, where Qiu *et al.* (2011) reported that codon usage bias in *A. thaliana* is much lower than in *D. melanogaster*. In addition, they found evidence that codon usage bias is weaker in selfers than in outcrosser.

In grasses, the codon usage bias has been mostly investigated in *O. sativa* and *Z. mays*. In *Z. mays*, Liu *et al.* (2010) reported that nucleotide composition and the level of gene expression were the main factors for codon usage bias while the variation in codon usage among genes is supposed to be due to mutational bias at the DNA level and natural selection acting on mRNA translation. In contrast, Wang and Hickey (2007) reported that in *O. sativa* the variation in codon usage bias is not related to selection acting on translation efficiency. Comparative studies of the codon bias usage in grasses done in the chloroplast of *B. distachyon*, *T. aestivum* and *H. vulgare* have shown that codons with `A` or `U` in the third position were preferred among those species (Sablok *et al.*, 2011). However, the rate of codon

usage and evolutionary patterns in the chloroplast differs from codon usage in nuclear genomes (Pfitzinger *et al.*, 1987). Those studies indicate that in grasses codon usage bias exists but is not completely understood.

Codon usage bias could explain why EDT from CDS with only synonymous sites were still more recent than EDTs derived from intergenic sequences were we removed putative methylation sites and indicates that also in *B. distachyon* codon usage bias exists. However, using another type of sequence, e.g. intergenic sequences can be used to compensate for the unknown codon usage bias in CDS sequences. Therefore, using several types of sequences can help to compensate for unknown influences on the molecular level when estimating divergence times.

4–4 Methods

The algorithms as well as the used strategies are described in more detail in the results section (page 92). Sequence alignments were done using `water` from the `EMBOSS` package version 6.3.1 (Rice *et al.*, 2000). If not stated otherwise, the following parameters were used: `-gapextend 0.1` for gap extension penalty and `-gapopen 30` for gap opening penalty. Alignments of CDS sequences were screened for synonymous sites with the `Perl` program `date_pair_protein`. Putative `C→T` methylated sites were removed using the `Perl` program `rmMethylPos.pl`. The start and end positions of LTR in the analyzed LTR-retrotransposons were defined with `DOTTER` version 3.1 (Sonnhammer and Durbin, 1995) by aligning the elements against itself.

Divergence Time Estimates

The divergence time estimations were calculated as described below. Let τ_i and τ_v be the number of transitions and transversion, respectively and *si* the sum of aligned sites in an sequence alignment. The fraction of nucleotides showing transversions is then $p = \frac{\tau_i}{si}$. Similarly, the fraction of nucleotides showing transitions is then $q = \frac{\tau_v}{si}$. Those terms where applied into the formulae from Kimura (1980) shown below to calculate the evolutionary distance per site ($k2p$) and the corresponding

standard error (SE_{k2p}).

$$k2p = -\frac{1}{2} \cdot \log\left((1 - 2p - q)\sqrt{1 - 2q}\right)$$
$$SE_{k2p} = \sqrt{\frac{1}{si}\left((a^2p + b^2q) - (ap + bq)^2\right)}$$

where a and b are

$$a = \frac{1}{1 - 2p - q}$$
$$b = \frac{1}{2} \cdot \log\left((1 - 2p - q) \cdot \sqrt{1 - 2q}\right)$$

The estimated divergence time (EDT) and corresponding standard error (SE_{EDT}) where than calculated as follows:

$$EDT = \frac{k2p}{2T}$$
$$SE_{EDT} = \frac{SE_{k2p}}{2T}$$

where T is the substitution rate. For our estimates we used $T = 1.3 \times 10^{-8} \frac{\frac{\text{substitutions}}{\text{per site}}}{\text{year}}$ (Ma and Bennetzen, 2004). The `Perl` programs developed for this study can be obtained as `git` repository from Jan P. Buchmann (`jbuchmann@botinst.uzh.ch`) or Dr. Thomas Wicker (`wicker@botinst.uzh.ch`)

General Discussion

5

A deluge of genomic plant sequences is expected in the near future, e.g. different subspecies from rice (OMAP, 2012), the barley genome (IBSC, 2012) and the 1001 Genomes Project from *Arabidopsis thaliana* (Cao *et al.*, 2011), just to name a few. The quality of those sequences will play an important role to identify and characterize new evolutionary mechanisms. The expected genomes will be mostly sequenced using next generation sequencing technologies (NGS), whose assemblies need sophisticated algorithms and software, or in case of close relatives, high quality reference genomes. While today *de novo* assemblies are feasible for genomes with a low amount of repetitive elements, e.g. *Brassica rapa* (Wang *et al.*, 2011), most plant genomes consist mainly of non-genic, repetitive sequences. In the case of grasses the amount of repetitive sequences can make up to 80%. The assembly of these sequences is, due to their repetitive nature, a very complex task. Many reads are excluded from the assembly as they consist entirely of repetitive sequences. In addition, repetitive sequences are sometimes collapsed into consensus sequences, whereby several reads are mapped to one location and therefore are actually missing at their original position. This means that genic regions are mostly assembled with high quality, but intergenic sequences contain lots of gaps or are found in unordered contigs. Therefore, the analysis of intergenic sequences which can constitute the majority of a plant genome has been very complicated or impossible due to low quality or missing sequence data, respectively.

The availability of high-quality genomic sequences from larger loci (e.g. *Adh*, Avramova *et al.* 1995; *Lr34*, Bossolini *et al.* 2007; *Glu-A3*, Wicker *et al.* 2003b)

allows detailed studies of the molecular mechanisms and composition of intergenic sequences. Analysis of the intergenic space depends on completely sequenced and long continuous sequences and is not feasible using strongly fragmented genome sequences. Therefore, no large and detailed study of the intergenic space in plants has been done yet.

We could show that a well annotated and assembled intergenic genome sequence offers an opportunity to identify mechanisms for genome evolution which could not have been identified using only sequences from genic regions.

5–1 Why Intergenic Sequence Quality and Annotation Matters

The annotation in the recently sequenced grass *Brachypodium distachyon* used a new approach. Genes and transposable elements (TEs) were annotated separately, reducing the usually frequent false annotation of TEs as genes. This can be observed by comparing numbers published by the Rice Genome Annotation Project (RGAP, `http://rice.plantbiology.msu.edu`). The *Oryza sativa* genome annotation first estimated a gene number between 32,000 and 50,000 and 4,814 TE sequences (Goff *et al.*, 2002). Newer releases of the RGAP regularly reduced the number of genes and increased the number of TEs (42,653 genes, 13,237 TE related sequences, Ouyang *et al.* (2007); 39,045 Non-TE loci and 49,066 TE loci, RGAP (2012)). It has to be noted that rice was the first completely sequenced grass genome and we have learned a lot since then, therefore the change in numbers is not surprising. However, TEs which are wrongly annotated as genes can still be found today.

The analysis of TEs is complicated. Several programs like `LTR_STRUC` (McCarthy and McDonald, 2003) or pipelines like `REPET` (Flutre *et al.*, 2011) have been developed to automate the analysis of TEs. They check for structural characteristics to identify one putative TE which then is used to identify its homologs in the genome. If several copies exist, it is classified as TE and a consensus from the copies is generated.

Most retrotransposons offer several characteristics which can be detected (e.g. Long Terminal Repeats (LTRs), primer binding sites and a polypurine tract, two Open Reading Frames (ORFs) and a length of 20 kilo base pair (kb) and more). In contrast, DNA-Transposons are much shorter, averaging 4–5 kb. Their only characteristics are Terminal Inverted Repeats (TIRs) and an ORF which encodes a transposase. The lack of common characteristics and their short length makes identification of DNA-Transposons more difficult.

In our comparative analysis between *B. distachyon* and *B. sylvaticum* we used an additional approach to find TEs. We examined the insertion and excision sites in both sequences for putative target site duplications (TSDs). Indeed, we identified several new, not yet characterized TEs, some of which also had a low copy number, e.g. below ten copies. Thus, comparison of insertions and excision of two closely related species is more sensitive than *de novo* predictions of TEs. However, this approach has two requirements: at least two closely related genomes are needed and the analysis of insertion sites involves manual work and is therefore very time consuming, especially on a genome wide level.

To analyze the upcoming flood of data, new software and tools need to be developed which ease the identification of new TEs and automatically analyze genomes and classify the, at least, obvious cases of insertions and deletions. The development and maintenance of reliable TE databases will improve the annotation and analysis of the genomes to come.

5–2 Analyzed Intergenic Sequences Reveal Molecular Mechanisms of Evolution

In our studies we showed that not only retrotransposons are responsible for the fast intergenic exchange, but the smaller DNA-Transposons also play a major role in shaping the intergenic landscape. We observed that the insertions of TEs,

independently of Class I or Class II, were precise, not altering the composition of the flanking sequences at the insertion site. In addition, most analyzed polymorphic sites between *B. sylvaticum* and *B. distachyon* harbored a DNA-Transposon in one or the other sequence. We found characteristic motifs for double-strand break (DSB) repair at the excision sites. Those models can explain the loss of sequence due to DSB repair by simple sequence annealing (SSA) or insertion of filler sequence due to repair by sequence dependent strand annealing (SDSA) (Puchta, 2005; Hartlerode and Scully, 2009). Our model that excision of DNA-Transposons causes those changes in sequence composition due to initiation of DSB repairs after excision is supported by a χ^2 test. The distribution of the deletions is not random but significantly linked to DNA-Transposon excision sites. illegitimate recombination (IR) also leads to a DSB. However, no mechanism have been identified yet. Our model offers an elegant and simple solution explaining most, if not all, IR signatures.

DNA-Transposons do not change the sequence composition due to their size but rather due to an indirect influence through DSB repair by the host cell machinery upon DNA-Transposon excision. In contrast to Class I elements, Class II elements introduce two DSBs within one transposition: the first at the excision and the second at the insertion site. As mentioned above, insertion do not alter the sequence composition. Nevertheless, Wicker *et al.* (2010) reported a case where the insertion of a *Mutator* element in *B. distachyon* led to a duplication of a gene which was copied and moved to the new location due to the DSB repair.

We could show that, depending on the used DSB repair mechanism, either sequence information is lost or non-homologous sequence information is introduced at the insertion site (Agmon *et al.*, 2009). In our dataset approximately 5% of the intergenic sequences were removed while 17% were exchanged. These findings demonstrate that intergenic sequences offer a possibility to discover evolutionary mechanism which would not have found using only genic regions. Further analysis

of intergenic regions will uncover yet unknown mechanisms or finding explanations for observations in genome evolution which could not be answered until today.

5–3 Divergence Time Estimates Are Complex

We estimated the divergence time between *B. distachyon* and *B. sylvaticum* by using three types of sequences: coding sequence (CDS), intergenic and LTR-retrotransposon. The comparison showed that deriving divergence estimates is difficult and several factors have to be considered. One is the amount of sequence data. Removing sites which are most likely under evolutionary pressure requires large datasets as using only assumed synonymous sites in CDS sequences removes approximately 80% of sequence information.

Estimated divergence times always represent a time range. This can be observed by using different sequences whose estimated divergence times were similar but not identical. However, by combining the different estimates we were able to define a lower limit for the estimated divergence time between *B. sylvaticum* and *B. distachyon*. Therefore, we propose that divergence estimates should be performed using different type of sequences. Using only synonymous sites in CDS sequences would not account for the codon usage bias. The inclusion of other types of sequences can compensate for influences which are not yet known. For example, intergenic sequences, which are not transcribed, can compensate for the codon usage bias. In addition, sequences from several different loci should be used. At the *All* locus we observed estimates which were not expected, i.e. the estimates derived from the intergenic dataset depicted a more recent divergence time than those derived from CDS sequences. An explanation could be the presence of deep paralogs. Deep paralogs are loci which had been duplicated already in the common ancestor and each of the descendants looses another copy of these loci. The divergence time estimation based on only such loci would result in estimates

for the deep paralogs but not for the actual species. Using several loci across the genome reduces this risk since it is very unlikely that all chosen loci represent deep paralogs.

The estimation of divergence times is based on two inputs, molecular data and substitution rates. The molecular data can be obtained and analyzed with high accuracy. We used two methods to reconstruct a phylogenetic tree based on *CACTA* transposase protein sequences. Both trees showed high similarity, supporting different models for molecular evolution. In contrast, substitution rates and the corresponding calibration using fossil records are very error prone.

In grasses, the first substitution rate for synonymous sites was published by Gaut *et al.* (1996) which used the *Adh* genes from rice and palm as molecular data and corresponding fossil divergence dates (Stebbins, 1981) for calibration of the substitution rate. This substitution rate was updated by Ma and Bennetzen (2004), which also used the *Adh* locus but from sorghum and maize instead. This updated substitution rate for synonymous sites is widely used in plant biology today and has not been updates since then. However, since the substitution rates have been determined in the lineage of grasses, they are quite reliable.

Graur and Martin (2004) showed that in mammals several divergence times estimates were actually not reliable due to the use of one single calibration which was extrapolated and stripped from error. Therefore, the authors propose to use several calibration points in each lineage. In addition, calibrations derived from one lineage should not be used for another, e.g. substitution rates from plants should not be used in estimation times for mammals. Therefore, divergence times can be subject to change. However, for most evolutionary analysis the change in absolute divergence times is negligible as long as the relative age is not changed (e.g. species A is older then species B) since most discovered evolutionary mechanisms are not based on absolute divergence times.

References

AGI (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.

Agmon, N., Pur, S., Liefshitz, B. and Kupiec, M. (2009) Analysis of repair mechanism choice during homologous recombination. *Nucleic Acids Res.*, **37**, 5081–5092.

Akashi, H. (1994) Synonymous codon usage in *Drosophila melanogaster* : natural selection and translational accuracy. *Genetics*, **136**, 927–935.

Alix, K., Joets, J., Ryder, C.D., Moore, J., Barker, G.C., Bailey, J.P., King, G.J. and Heslop-Harrison, J.S.P. (2008) The CACTA transposon *Bot1* played a major role in *Brassica* genome divergence and gene proliferation. *Plant J.*, **56**, 1030–1044.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **17**, 3389–3402.

Argout, X., Salse, J., Aury, J.M., Guiltinan, M.J., Droc, G., Gouzy, J., Allegre, M., Chaparro, C., Legavre, T., Maximova, S.N. *et al.* (2011) The genome of *Theobroma cacao*. *Nat. Genet.*, **43**, 101–108.

Athma, P. and Peterson, T. (1991) *Ac* induces homologous recombination at the maize *P* locus. *Genetics*, **128**, 163–173.

Avramova, Z., SanMiguel, P., Georgieva, E. and Bennetzen, J.L. (1995) Matrix attachment regions and transcribed sequences within a long chromosomal continuum containing maize *Adh1*. *Plant Cell*, **7**, 1667–1680.

Ayala, F.J., Rzhetsky, A. and Ayala, F.J. (1998) Origin of the metazoan phyla: molecular clocks confirm paleontological estimates. *Proc. Natl Acad. Sci. USA*, **95**, 606–611.

Barakat, A., Matassi, G. and Bernardi, G. (1998) Distribution of genes in the genome of *Arabidopsis thaliana* and its implications for the genome organization of plants. *Proc. Natl Acad. Sci. USA*, **95**, 10044–10049.

Bennett, M.D. and Leitch, I.J. (1995) Nuclear DNA Amounts in Angiosperms. *Ann.Bot-London*, **76**, 113–176.

Bennetzen, J.L. (2000) Transposable element contributions to plant gene and genome evolution. *Plant Mol. Biol.*, **42**, 251–269.

Bennetzen, J.L., Schrick, K., Springer, P.S., Brown, W.E. and SanMiguel, P. (1994) Active maize genes are unmodified and flanked by diverse classes of modified, highly repetitive DNA. *Genome*, **37**, 565–576.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Wheeler, D.L. (2008) GenBank. *Nucleic Acids Res.*, **36(189.0)**, D25–D30.

Bonierbale, M.W., Plaisted, R.L. and Tanksley, S.D. (1988) RFLP Maps Based on a Common Set of Clones Reveal Modes of Chromosomal Evolution in Potato and Tomato. *Genetics*, **120**, 1095–1103.

Bossolini, E., Wicker, T., Knobel, P.A. and Keller, B. (2007) Comparison of orthologous loci from small grass genomes *Brachypodium* and rice: implications for wheat genomics and grass genome annotation. *Plant J.*, **49**, 704–717.

Boutabout, M., Wilhelm, M. and Wilhelm, F.X. (2001) DNA synthesis fidelity by the reverse transcriptase of the yeast retrotransposon Ty1. *Nucleic Acids Res.*, **29**, 2217–2222.

Braquart, C., Royer, V. and Bouhin, H. (1999) DEC: a new miniature inverted-repeat transposable element from the genome of the beetle *Tenebrio molitor*. *Insect Mol. Biol.*, **8**, 571–574.

Brillet, B., Benjamin, B., Bigot, Y., Yves, B., Augé-Gouillou, C. and Corinne, A.G. (2007) Assembly of the *Tc1* and *mariner* transposition initiation complexes depends on the origins of their transposase DNA binding domains. *Genetica*, **130**, 105–120.

Britten, R.J. (1986) Rates of DNA sequence evolution differ between taxonomic groups. *Science*, **231**, 1393–1398.

Bryan, G., Garza, D. and Hartl, D. (1990) Insertion and excision of the transposable element *mariner* in Drosophila. *Genetics*, **125**, 103–114.

Buchmann, J.P., Keller, B. and Wicker, T. (in press) *Transposons in cereals: shaping genomes and driving their evolution.* Springer.

Buchmann, J.P., Matsumoto, T., Stein, N., Keller, B. and Wicker, T. (2012) Inter-species sequence comparison of *Brachypodium* reveals how transposon activity corrodes genome colinearity. *Plant J.*, Accepted Article.

Bureau, T.E. and Wessler, S.R. (1992) *Tourist*: a large family of small inverted repeat elements frequently associated with maize genes. *Plant Cell*, **4**, 1283–1294.

Bureau, T.E. and Wessler, S.R. (1994a) Mobile inverted-repeat elements of the *Tourist* family are associated with the genes of many cereal grasses. *Proc. Natl Acad. Sci. USA*, **91**, 1411–1415.

Bureau, T.E. and Wessler, S.R. (1994b) *Stowaway*: a new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants. *Plant Cell*, **6**, 907–916.

Cao, J., Schneeberger, K., Ossowski, S., Günther, T., Bender, S., Fitz, J., Koenig, D., Lanz, C., Stegle, O., Lippert, C. *et al.* (2011) Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nat Genet*, **43**, 956–963.

Casacuberta, J.M. and Santiago, N. (2003) Plant LTR-retrotransposons and MITEs: control of transposition and impact on the evolution of plant genes and genomes. *Gene*, **311**, 1–11.

Chain, P.S.G., Grafham, D.V., Fulton, R.S., Fitzgerald, M.G., Hostetler, J., Muzny, D., Ali, J., Birren, B., Bruce, D.C., Buhay, C. *et al.* (2009) Genomics. Genome project standards in a new era of sequencing. *Science*, **326**, 236–237.

Chamary, J.V., Parmley, J.L. and Hurst, L.D. (2006) Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat Rev Genet*, **7**, 98–108.

Chao, S., Sharp, P.J., Worland, A.J., Warham, E.J., Koebner, R.M.D. and Gale, M.D. (1989) RFLP-based genetic maps of wheat homoeologous group 7 chromosomes. *Theor. Appl. Genet.*, **78**, 495–504. 10.1007/BF00290833.

Chao, S., Sharp, P. and Gale, M. (1988) *Proc. 7th Int. Wheat Genet. Symp.* IPSR, Cambridge Laboratory, Cambridge.

Chen, J.M., Cooper, D.N., Chuzhanova, N., Férec, C. and Patrinos, G.P. (2007) Gene conversion: mechanisms, evolution and human disease. *Nat Rev Genet*, **8**, 762–775.

Chopra, S., Brendel, V., Zhang, J., Axtell, J.D. and Peterson, T. (1999) Molecular characterization of a mutable pigmentation phenotype and isolation of the first active transposable element from *Sorghum bicolor*. *Proc. Natl Acad. Sci. USA*, **96**, 15330–15335.

Daghlian, C.P. (1981) A Review of the fossil record of monocotyledons. *Botanical Review*, **47**, 517–555.

Dawson, A. and Finnegan, D.J. (2003) Excision of the *Drosophila* mariner transposon Mos1. Comparison with bacterial transposition and V(D)J recombination. *Mol. Cell*, **11**, 225–235.

Devos, K.M. (2005) Updating the 'Crop Circle'. *Curr. Opin. Plant Biol.*, **8**, 155–162.

Devos, K.M., Atkinson, M.D., Chinoy, C.N., Francis, H.A., Harcourt, R.L., Koebner, R.M.D., Liu, C.J., Masojć, P., Xie, D.X. and Gale, M.D. (1993) Chromosomal rearrangements in the rye genome relative to that of wheat. *Theor. Appl. Genet.*, **85**, 673–680. 10.1007/BF00225004.

Devos, K.M., Brown, J.K.M. and Bennetzen, J.L. (2002) Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. *Genome Res.*, **12**, 1075–1079.

Doolittle, W.F. and Sapienza, C. (1980) Selfish genes, the phenotype paradigm and genome evolution. *Nature*, **284**, 601–603.

Duvall, M.R., Learn, G.H., Eguiarte, L.E. and Clegg, M.T. (1993) Phylogenetic analysis of *rbcL* sequences identifies *Acorus calamus* as the primal extant monocotyledon. *Proc. Natl Acad. Sci. USA*, **90**, 4641–4644.

Faris, J.D., Zhang, Z., Fellers, J.P. and Gill, B.S. (2008) Micro-colinearity between rice, *Brachypodium*, and *Triticum monococcum* at the wheat domestication locus *Q. Funct. Integr. Genomics*, **8**, 149–164.

Fedoroff, N., Wessler, S. and Shure, M. (1983) Isolation of the transposable maize controlling elements *Ac* and *Ds*. *Cell*, **35**, 235–242.

Felsenstein, J. (2005) PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle. `http://evolution.genetics.washington.edu/phylip.html`.

Feschotte, C. and Mouchès, C. (2000) Recent amplification of miniature inverted-repeat transposable elements in the vector mosquito *Culex pipiens*: characterization of the *Mimo* family. *Gene*, **250**, 109–116.

Feschotte, C., Swamy, L. and Wessler, S.R. (2003) Genome-wide analysis of *mariner*-like transposable elements in rice reveals complex relationships with *stowaway* miniature inverted repeat transposable elements (MITEs). *Genetics*, **163**, 747–758.

Fischer, S.E., Wienholds, E. and Plasterk, R.H. (2001) Regulated transposition of a fish transposon in the mouse germ line. *Proc. Natl Acad. Sci. USA*, **98**, 6759–6764.

Flavell, R.B., Bennett, M.D., Smith, J.B. and Smith, D.B. (1974) Genome size and the proportion of repeated nucleotide sequence DNA in plants. *Biochem. Genet.*, **12**, 257–269.

Flutre, T., Duprat, E., Feuillet, C. and Quesneville, H. (2011) Considering transposable element diversification in *de novo* annotation approaches. *PLoS One*, **6**, e16526.

Foote, T.N., Griffiths, S., Allouis, S. and Moore, G. (2004) Construction and analysis of a BAC library in the grass *Brachypodium sylvaticum*: its use as a tool to bridge the gap between rice and wheat in elucidating gene content. *Funct. Integr. Genomics*, **4**, 26–33.

Freeling, M. and Subramaniam, S. (2009) Conserved noncoding sequences (CNSs) in higher plants. *Curr. Opin. Plant Biol.*, **12**, 126–132.

Frey, M., Reinecke, J., Grant, S., Saedler, H. and Gierl, A. (1990) Excision of the En/Spm transposable element of *Zea mays* requires two element-encoded proteins. *EMBO J.*, **9**, 4037–4044.

Gale, M.D. and Devos, K.M. (1998) Comparative genetics in the grasses. *Proc. Natl. Acad. Sci. USA*, **95**, 1971–1974.

Gallego, F., Feuillet, C., Messmer, M., Penger, A., Graner, A., Yano, M., Sasaki, T. and Keller, B. (1998) Comparative mapping of the two wheat leaf rust resistance loci *Lr1* and *Lr10* in rice and barley. *Genome*, **41**, 328–336.

Gaut, B.S., Morton, B.R., McCaig, B.C. and Clegg, M.T. (1996) Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene *Adh* parallel rate differences at the plastid gene *rbcL*. *Proc. Natl. Acad. Sci. USA*, **93**, 10274–10279.

Gill, K.S., Gill, B.S., Endo, T.R. and Taylor, T. (1996) Identification and high-density mapping of gene-rich regions in chromosome group 1 of wheat. *Genetics*, **144**, 1883–1891.

Goff, S.A., Ricke, D., Lan, T.H., Presting, G., Wang, R., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P., Varma, H. *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science*, **296**, 92–100.

Goloubinoff, P., Pääbo, S. and Wilson, A.C. (1993) Evolution of maize inferred from sequence diversity of an *Adh2* gene segment from archaeological specimens. *Proc. Natl Acad. Sci. USA*, **90**, 1997–2001.

Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N. *et al.* (2012) Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.*, **40**, D1178–D1186. Accessed 2012.05.18.

Gorbunova, V. and Levy, A.A. (1999) How plants make ends meet: DNA double–strand break repair. *Trends Plant Sci.*, **4**, 263–269.

Grandbastien, M.A., Audeon, C., Bonnivard, E., Casacuberta, J.M., Chalhoub, B., Costa, A.P.P., Le, Q.H., Melayah, D., Petit, M., Poncet, C. *et al.* (2005) Stress activation and genomic impact of Tnt1 retrotransposons in Solanaceae. *Cytogenet. Genome Res.*, **110**, 229–241.

Graur, D. and Martin, W. (2004) Reading the entrails of chickens: molecular timescales of evolution and the illusion of precision. *Trends Genet.*, **20**, 80–86.

Greco, R., Ouwerkerk, P.B.F. and Pereira, A. (2005) Suppression of an atypically spliced rice CACTA transposon transcript in transgenic plants. *Genetics*, **169**, 2383–2387.

Gregory, T.R., Nicol, J.A., Tamm, H., Kullman, B., Kullman, K., Leitch, I.J., Murray, B.G., Kapraun, D.F., Greilhuber, J. and Bennett, M.D. (2007) Eukaryotic genome size databases. *Nucleic Acids Res.*, **35**, D332–D338.

Greider, C.W. and Blackburn, E.H. (1985) Identification of a specific telomere terminal transferase activity in Tetrahymena extracts. *Cell*, **43**, 405–413.

Greider, C.W. and Blackburn, E.H. (1987) The telomere terminal transferase of Tetrahymena is a ribonucleoprotein enzyme with two kinds of primer specificity. *Cell*, **51**, 887–898.

Griffiths, S., Sharp, R., Foote, T.N., Bertin, I., Wanous, M., Reader, S., Colas, I. and Moore, G. (2006) Molecular characterization of *Ph1* as a major chromosome pairing locus in polyploid wheat. *Nature*, **439**, 749–752.

Guyot, R. and Keller, B. (2004) Ancestral genome duplication in rice. *Genome*, **47**, 610–614.

Haber, J.E., Ira, G., Malkova, A. and Sugawara, N. (2004) Repairing a double-strand chromosome break by homologous recombination: revisiting Robin Holliday's model. *Philos Trans R Soc Lond B Biol Sci*, **359**, 79–86.

Hartlerode, A.J. and Scully, R. (2009) Mechanisms of double-strand break repair in somatic mammalian cells. *Biochem. J.*, **423**, 157–168.

Holmquist, G.P. and Ashley, T. (2006) Chromosome organization and chromatin modification: influence on genome function and evolution. *Cytogenet. Genome Res.*, **114**, 96–125.

Hu, T.T., Pattyn, P., Bakker, E.G., Cao, J., Cheng, J.F., Clark, R.M., Fahlgren, N., Fawcett, J.A., Grimwood, J., Gundlach, H. *et al.* (2011) The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nat. Genet.*, **43**, 476–481.

Huang, X., Lu, G., Zhao, Q., Liu, X. and Han, B. (2008) Genome-wide analysis of transposon insertion polymorphisms reveals intraspecific variation in cultivated rice. *Plant Physiol.*, **148**, 25–40.

Hurwitz, B.L., Kudrna, D., Yu, Y., Sebastian, A., Zuccolo, A., Jackson, S.A., Ware, D., Wing, R.A. and Stein, L. (2010) Rice structural variation: a comparative analysis of structural variation between rice and three of its closest relatives in the genus Oryza. *Plant J.*, **63**, 990–1003.

IBI (2010) Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature*, **463**, 763–768.

IBSC (2012) The International Barley Sequencing Consortium. `http://www.public.iastate.edu/~imagefpc/IBSC%20Webpage/IBSC%20Template-home.html`. Online accessed: 21.07.2012.

Inagaki, Y., Hisatomi, Y., Suzuki, T., Kasahara, K. and Iida, S. (1994) Isolation of a *Suppressor-mutator/Enhancer*-like transposable element, *Tpn1*, from Japanese morning glory bearing variegated flowers. *Plant Cell*, **6**, 375–383.

IRGSP (2005) The map-based sequence of the rice genome. *Nature*, **436**, 793–800.

Isidore, E., Scherrer, B., Chalhoub, B., Feuillet, C. and Keller, B. (2005) Ancient haplotypes resulting from extensive molecular rearrangements in the wheat A genome have been maintained in species of three different ploidy levels. *Genome Res.*, **15**, 526–536.

Izsvák, Z., Ivics, Z., Shimoda, N., Mohn, D., Okamoto, H. and Hackett, P.B. (1999) Short inverted-repeat transposable elements in teleost fish and implications for a mechanism of their amplification. *J. Mol. Evol.*, **48**, 13–21.

Jaillon, O., Aury, J.M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., Choisne, N., Aubourg, S., Vitulo, N., Jubin, C. *et al.* (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, **449**, 463–467.

Kalendar, R., Tanskanen, J., Immonen, S., Nevo, E. and Schulman, A.H. (2000) Genome evolution of wild barley (*Hordeum spontaneum*) by *BARE*-1 retrotransposon dynamics in response to sharp microclimatic divergence. *Proc. Natl Acad. Sci. USA*, **97**, 6603–6607.

Kalendar, R., Tanskanen, J., Chang, W., Antonius, K., Sela, H., Peleg, O. and Schulman, A.H. (2008) *Cassandra* retrotransposons carry independently transcribed 5S RNA. *Proc. Natl Acad. Sci. USA*, **105**, 5833–5838.

Kalendar, R., Vicient, C.M., Peleg, O., Anamthawat-Jonsson, K., Bolshoy, A. and Schulman, A.H. (2004) Large retrotransposon derivatives: abundant, conserved but nonautonomous retroelements of barley and related genomes. *Genetics*, **166**, 1437–1450.

Kapitonov, V.V. and Jurka, J. (2005) RAG1 core and V(D)J recombination signal sequences were derived from *Transib* transposons. *PLoS Biol.*, **3**, e181.

Kimura, M. (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.*, **16**, 111–120.

Kouzarides, T. (2007) SnapShot: Histone-modifying enzymes. *Cell*, **131**, 822.

Krishan, A., Dandekar, P., Nathan, N., Hamelik, R., Miller, C. and Shaw, J. (2005) DNA index, genome size, and electronic nuclear volume of vertebrates from the Miami Metro Zoo. *Cytometry A.*, **65**, 26–34.

Kumar, A. and Bennetzen, J.L. (1999) Plant retrotransposons. *Annu. Rev. Genet.*, **33**, 479–532.

Künzel, G., Korzun, L. and Meister, A. (2000) Cytologically integrated physical restriction fragment length polymorphism maps for the barley genome based on translocation breakpoints. *Genetics*, **154**, 397–412.

Lafay, B., Atherton, J.C. and Sharp, P.M. (2000) Absence of translationally selected synonymous codon usage bias in *Helicobacter pylori*. *Microbiology*, **146 (Pt 4)**, 851–860.

Lafay, B., Lloyd, A.T., McLean, M.J., Devine, K.M., Sharp, P.M. and Wolfe, K.H. (1999) Proteome composition and codon usage in spirochaetes: species-specific and DNA strand-specific mutational biases. *Nucleic Acids Res*, **27**, 1642–1649.

Langdon, T., Jenkins, G., Hasterok, R., Jones, R.N. and King, I.P. (2003) A high-copy-number CACTA family transposon in temperate grasses and cereals. *Genetics*, **163**, 1097–1108.

Levis, R.W. (1989) Viable deletions of a telomere from a Drosophila chromosome. *Cell*, **58**, 791–801.

Lewin, B. (1997) *Genes VI*. Oxford University Press, Inc. New York.

Lisch, D. (2009) Epigenetic regulation of transposable elements in plants. *Annu. Rev. Plant Biol.*, **60**, 43–66.

Liu, H., He, R., Zhang, H., Huang, Y., Tian, M. and Zhang, J. (2010) Analysis of synonymous codon usage in *Zea mays*. *Mol Biol Rep*, **37**, 677–684.

Long, M., de Souza, S.J., Rosenberg, C. and Gilbert, W. (1996) Exon shuffling and the origin of the mitochondrial targeting function in plant cytochrome c1 precursor. *Proc. Natl Acad. Sci. USA*, **93**, 7727–7731.

Lowe, B., Mathern, J. and Hake, S. (1992) Active *Mutator* elements suppress the knotted phenotype and increase recombination at the *Kn1-O* tandem duplication. *Genetics*, **132**, 813–822.

Lu, C., Chen, J., Zhang, Y., Hu, Q., Su, W. and Kuang, H. (2012) Miniature Inverted-Repeat Transposable Elements (MITEs) have been accumulated through amplification bursts and play important roles in gene expression and species diversity in *Oryza sativa*. *Mol. Biol. Evol.*, **29**, 1005–1017.

Ma, J. and Bennetzen, J.L. (2004) Rapid recent growth and divergence of rice nuclear genomes. *Proc. Natl Acad. Sci. USA*, **101**, 12404–12410.

Ma, J., Devos, K.M. and Bennetzen, J.L. (2004) Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. *Genome Res.*, **14**, 860–869.

Manninen, I. and Schulman, A.H. (1993) *BARE-1*, a copia-like retroelement in barley (*Hordeum vulgare* L.). *Plant Mol. Biol.*, **22**, 829–846.

Marshall, C.R. (1990) The fossil record and estimating divergence times between lineages: Maximum divergene times and the importance of reliable phylogenies. *J. Mol. Evol.*, **30**, 400–408. 10.1007/BF02101112.

Mason-Gamer, R.J. (2007) Multiple homoplasious insertions and deletions of a Triticeae (Poaceae) DNA transposon: a phylogenetic perspective. *BMC Evol. Biol.*, **7**, 92.

Masson, P., Strem, M. and Fedoroff, N. (1991) The *tnpA* and *tnpD* gene products of the *Spm* element are required for transposition in tobacco. *Plant Cell*, **3**, 73–85.

Matzke, M., Kanno, T., Daxinger, L., Huettel, B. and Matzke, A.J.M. (2009) RNA-mediated chromatin-based silencing in plants. *Curr. Opin. Cell Biol.*, **21**, 367–376.

Mayer, K.F.X., Martis, M., Hedley, P.E., Simková, H., Liu, H., Morris, J.A., Steuernagel, B., Taudien, S., Roessner, S., Gundlach, H. *et al.* (2011) Unlocking the barley genome by chromosomal and comparative genomics. *Plant Cell*, **23**, 1249–1263.

McCarthy, E.M. and McDonald, J.F. (2003) LTR_STRUC: a novel search and identification program for LTR retrotransposons. *Bioinformatics*, **19**, 362–367.

McClintock, B. (1950) The origin and behavior of mutable loci in maize. *Proc. Natl Acad. Sci. USA*, **36**, 344–355.

McClintock, B. (1984) The significance of responses of the genome to challenge. *Science*, **226**, 792–801.

Menzel, G., Krebs, C., Diez, M., Holtgräwe, D., Weisshaar, B., Minoche, A.E., Dohm, J.C., Himmelbauer, H. and Schmidt, T. (2012) Survey of sugar beet (*Beta vulgaris* L.) *hAT transposons* and MITE-like *hATpin* derivatives. *Plant Mol. Biol.*, **78**, 393–405.

Miller, J.T., Dong, F., Jackson, S.A., Song, J. and Jiang, J. (1998) Retrotransposon-related DNA sequences in the centromeres of grass chromosomes. *Genetics*, **150**, 1615–1623.

Miura, A., Yonebayashi, S., Watanabe, K., Toyama, T., Shimada, H. and Kakutani, T. (2001) Mobilization of transposons by a mutation abolishing full DNA methylation in *Arabidopsis*. *Nature*, **411**, 212–214.

Mizuuchi, K. (1997) Polynucleotidyl transfer reactions in site-specific DNA recombination. *Genes Cells*, **2**, 1–12.

Moore, G., Devos, K.M., Wang, Z. and Gale, M.D. (1995a) Cereal genome evolution. Grasses, line up and form a circle. *Curr. Biol.*, **5**, 737–739.

Moore, G., Foote, T., Helentjaris, T., Devos, K., Kurata, N. and Gale, M. (1995b) Was there a single ancestral cereal chromosome? *Trends Genet.*, **11**, 81–82.

Moore, G.E. (1965) Cramming more components onto integrated circuits. *Electronics*, **38**.

Nacken, W.K., Piotrowiak, R., Saedler, H. and Sommer, H. (1991) The transposable element *Tam1* from *Antirrhinum majus* shows structural homology to the maize transposon *En/Spm* and has no sequence specificity of insertion. *Mol. Gen. Genet.*, **228**, 201–208.

Nagaki, K., Neumann, P., Zhang, D., Ouyang, S., Buell, C.R., Cheng, Z. and Jiang, J. (2005) Structure, divergence, and distribution of the CRR centromeric retrotransposon family in rice. *Mol. Biol. Evol.*, **22**, 845–855.

Neumann, P., Navrátilová, A., Koblížková, A., Kejnovský, E., Hřibová, E., Hobza, R., Widmer, A., Doležel, J. and Macas, J. (2011) Plant centromeric retrotransposons: a structural and cytogenetic perspective. *Mob. DNA*, **2**, 4.

Oettinger, M.A., Schatz, D.G., Gorka, C. and Baltimore, D. (1990) RAG-1 and RAG-2, adjacent genes that synergistically activate V(D)J recombination. *Science*, **248**, 1517–1523.

Ohno, S. (1972) So much "junk" DNA in our genome. *Brookhaven Symp. Biol.*, **23**, 366–370.

OMAP (2012) The Oryza Map Alignment Project. `http://www.omap.org/index.html`. Online accessed: 21.07.2012.

Oosumi, T., Garlick, B. and Belknap, W.R. (1996) Identification of putative nonautonomous transposable elements associated with several transposon families in *Caenorhabditis elegans*. *J. Mol. Evol.*, **43**, 11–18.

Ouyang, S., Zhu, W., Hamilton, J., Lin, H., Campbell, M., Childs, K., Thibaud-Nissen, F., Malek, R.L., Lee, Y., Zheng, L. *et al.* (2007) The TIGR Rice Genome Annotation Resource: improvements and new features. *Nucleic Acids Res.*, **35**, D883–D887.

Ozeki, Y., Davies, E. and Takeda, J. (1997) Somatic variation during long term subculturing of plant cells caused by insertion of a transposable element in a phenylalanine ammonia-lyase (PAL) gene. *Mol. Gen. Genetics*, **254**, 407–416. 10.1007/s004380050433.

Pardue, M.L. and DeBaryshe, P.G. (2011) Retrotransposons that maintain chromosome ends. *Proc. Natl Acad. Sci. USA*, **108**, 20317–20324.

Pardue, M.L. and Debaryshe, P. (2011) Adapting to life at the end of the line: How Drosophila telomeric retrotransposons cope with their job. *Mob. Genet. Elements*, **1**, 128–134.

Paterson, A.H., Bowers, J.E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., Haberer, G., Hellsten, U., Mitros, T., Poliakov, A. *et al.* (2009) The *Sorghum bicolor* genome and the diversification of grasses. *Nature*, **457**, 551–556.

Pereira, A., Cuypers, H., Gierl, A., Schwarz-Sommer, Z. and Saedler, H. (1986) Molecular analysis of the En/Spm transposable element system of *Zea mays*. *EMBO J.*, **5**, 835–841.

Pereira, V. (2004) Insertion bias and purifying selection of retrotransposons in the *Arabidopsis thaliana* genome. *Genome Biol.*, **5**, R79.

Petrov, D.A. (2001) Evolution of genome size: new approaches to an old problem. *Trends Genet.*, **17**, 23–28.

Pfitzinger, H., Guillemaut, P., Weil, J.H. and Pillay, D.T. (1987) Adjustment of the tRNA population to the codon usage in chloroplasts. *Nucleic Acids Res*, **15**, 1377–1386.

Preston, B.D. (1996) Error-prone retrotransposition: rime of the ancient mutators. *Proc. Natl Acad. Sci. USA*, **93**, 7427–7431.

Puchta, H. (2005) The repair of double-strand breaks in plants: mechanisms and consequences for genome evolution. *J. Exp. Bot.*, **56**, 1–14.

Qiu, S., Zeng, K., Slotte, T., Wright, S. and Charlesworth, D. (2011) Reduced efficacy of natural selection on codon usage bias in selfing *Arabidopsis* and *Capsella* species. *Genome Biol Evol*, **3**, 868–880.

Ramallo, E., Kalendar, R., Schulman, A.H. and Martínez-Izquierdo, J.A. (2008) *Reme1*, a *Copia* retrotransposon in melon, is transcriptionally induced by UV light. *Plant Mol. Biol.*, **66**, 137–150.

Rensing, S.A., Lang, D., Zimmer, A.D., Terry, A., Salamov, A., Shapiro, H., Nishiyama, T., Perroud, P.F., Lindquist, E.A., Kamisugi, Y. *et al.* (2008) The *Physcomitrella* genome reveals evolutionary insights into the conquest of land by plants. *Science*, **319**, 64–69.

RGAP (2012) The Rice Genome Annotation Project. `http://rice.plantbiology.msu.edu`. Online; accessed 21.07.2012.

Rice, P., Longden, I. and Bleasby, A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.*, **16**, 276–277. http://emboss.sourceforge.net.

Rice, P.A. and Baker, T.A. (2001) Comparative architecture of transposase and integrase complexes. *Nat. Struct. Biol.*, **8**, 302–307.

Richardson, J.M., Colloms, S.D., Finnegan, D.J. and Walkinshaw, M.D. (2009) Molecular architecture of the Mos1 paired-end complex: the structural basis of DNA transposition in a eukaryote. *Cell*, **138**, 1096–1108.

Robert, V. and Bessereau, J.L. (2007) Targeted engineering of the *Caenorhabditis elegans* genome following *Mos1*-triggered chromosomal breaks. *EMBO J.*, **26**, 170–183.

Sablok, G., Nayak, K., Vazquez, F. and Tatarinova, T. (2011) Synonymous codon usage, GC_3, and evolutionary patterns across plastomes of three pooid model species: emerging grass genome models for monocots. *Molecular Biotechnology*, **49**, 116–128. 10.1007/s12033-011-9383-9.

Sabot, F. and Schulman, A.H. (2006) Parasitism and the retrotransposon life cycle in plants: a hitchhiker's guide to the genome. *Heredity (Edinb)*, **97**, 381–388.

Sabot, F., Sourdille, P., Chantret, N. and Bernard, M. (2006) Morgane, a new LTR retrotransposon group, and its subfamilies in wheats. *Genetica*, **128**, 439–447.

Sanderson, M.J. and Doyle, J.A. (2001) Sources of error and confidence intervals in estimating the age of angiosperms from *rbcL* and 18S rDNA data. *Am. J. Bot.*, **88**, 1499–1516.

SanMiguel, P., Gaut, B.S., Tikhonov, A., Nakajima, Y. and Bennetzen, J.L. (1998) The paleontology of intergene retrotransposons of maize. *Nat. Genet.*, **20**, 43–45.

SanMiguel, P.J., Ramakrishna, W., Bennetzen, J.L., Busso, C.S. and Dubcovsky, J. (2002) Transposable elements, genes and recombination in a 215-kb contig from wheat chromosome $5A^m$. *Funct. Integr. Genomics*, **2**, 70–80.

Scherrer, B., Isidore, E., Klein, P., Kim, J., Bellec, A., Chalhoub, B., Keller, B. and Feuillet, C. (2005) Large intraspecific haplotype variability at the *Rph7* locus results from rapid and recent divergence in the barley genome. *Plant Cell*, **17**, 361–374.

Schmidt, H.A., Strimmer, K., Vingron, M. and von Haeseler, A. (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics*, **18**, 502–504.

Schmidt, H.A. and von Haeseler, A. (2009) *Phylogenetic interference using maximum likelihood methods*, chapter 6. Cambridge University Press, 2nd edition.

Schmutz, J., Cannon, S.B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D.L., Song, Q., Thelen, J.J., Cheng, J. *et al.* (2010) Genome sequence of the palaeopolyploid soybean. *Nature*, **463**, 178–183.

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T.A. *et al.* (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science*, **326**, 1112–1115.

Schulman, A.H. and Kalendar, R. (2005) A movable feast: diverse retrotransposons and their contribution to barley genome dynamics. *Cytogenet. Genome Res.*, **110**, 598–605.

Schulman, A.H., Flavell, A.J. and Ellis, T.H.N. (2004) The application of LTR retrotransposons as molecular markers in plants. *Methods Mol. Biol.*, **260**, 145–173.

Sinzelle, L., Kapitonov, V.V., Grzela, D.P., Jursch, T., Jurka, J., Izsvák, Z. and Ivics, Z. (2008) Transposition of a reconstructed Harbinger element in human cells and functional homology with two transposon-derived cellular genes. *Proc. Natl Acad. Sci. USA*, **105**, 4715–4720.

Slightom, J.L., Blechl, A.E. and Smithies, O. (1980) Human fetal ^{G}gamma- and ^{A} gamma-globin genes: complete nucleotide sequences suggest that DNA can be exchanged between these duplicated genes. *Cell*, **21**, 627–638.

Smit, A.F. and Riggs, A.D. (1996) Tiggers and DNA transposon fossils in the human genome. *Proc. Natl Acad. Sci. USA*, **93**, 1443–1448.

Snowden, K.C. and Napoli, C.A. (1998) *Psl*: a novel *Spm*-like transposable element from *Petunia hybrida*. *Plant J.*, **14**, 43–54.

Sonnhammer, E.L. and Durbin, R. (1995) A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene*, **167**, GC1–G10.

Stebbins, G.L. (1981) Coevolution of Grasses and Herbivores. *Annals of the Missouri Botanical Garden*, **68**, 75–86.

Suoniemi, A., Tanskanen, J., Pentikäinen, O., Johnson, M.S. and Schulman, A.H. (1998) The core domain of retrotransposon integrase in *Hordeum*: predicted structure and evolution. *Mol. Biol. Evol.*, **15**, 1135–1144.

Szostak, J.W., Orr-Weaver, T.L., Rothstein, R.J. and Stahl, F.W. (1983) The double-strand-break repair model for recombination. *Cell*, **33**, 25–35.

Thompson, J.D.D.H. and Gibson, T. (1994) Clustal W, improvig the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673 – 4680.

Tonegawa, S. (1983) Somatic generation of antibody diversity. *Nature*, **302**, 575–581.

Tsukahara, S., Kobayashi, A., Kawabe, A., Mathieu, O., Miura, A. and Kakutani, T. (2009) Bursts of retrotransposition reproduced in *Arabidopsis*. *Nature*, **461**, 423–426.

Tu, Z. (1997) Three novel families of miniature inverted-repeat transposable elements are associated with genes of the yellow fever mosquito, *Aedes aegypti*. *Proc. Natl Acad. Sci. USA*, **94**, 7475–7480.

Ueki, N. and Nishii, I. (2008) *Idaten* is a new cold-inducible transposon of *Volvox carteri* that can be used for tagging developmentally important genes. *Genetics*, **180**, 1343–1353.

Unsal, K. and Morgan, G.T. (1995) A novel group of families of short interspersed repetitive elements (SINEs) in *Xenopus*: evidence of a specific target site for DNA-mediated transposition of inverted-repeat SINEs. *J. Mol. Biol.*, **248**, 812–823.

Vicient, C.M., Kalendar, R., Anamthawat-Jónsson, K. and Schulman, A.H. (1999) Structure, functionality, and evolution of the *BARE*-1 retrotransposon of barley. *Genetica*, **107**, 53–63.

Vitte, C. and Panaud, O. (2005) LTR retrotransposons and flowering plant genome size: emergence of the increase/decrease model. *Cytogenet. Genome Res.*, **110**, 91–107.

Vu, G.T.H., Wicker, T., Buchmann, J.P., Chandler, P.M., Matsumoto, T., Graner, A. and Stein, N. (2010) Fine mapping and syntenic integration of the semi-dwarfing gene *sdw3* of barley. *Funct. Integr. Genomics*, **10**, 509–521.

Wang, H.C. and Hickey, D.A. (2007) Rapid divergence of codon usage patterns within the rice genome. *BMC Evol Biol*, **7 Suppl 1**, S6.

Wang, X., Wang, H., Wang, J., Sun, R., Wu, J., Liu, S., Bai, Y., Mun, J.H., Bancroft, I., Cheng, F. *et al.* (2011) The genome of the mesopolyploid crop species *Brassica rapa*. *Nat. Genet.*, **43**, 1035–1039.

Watson, J.D. (1972) Origin of concatemeric T7 DNA. *Nat. New Biol.*, **239**, 197–201.

Weil, C.F. and Kunze, R. (2000) Transposition of maize *Ac/Ds* transposable elements in the yeast *Saccharomyces cerevisiae*. *Nat. Genet.*, **26**, 187–190.

Weiner, A.M., Deininger, P.L. and Efstratiadis, A. (1986) Nonviral retroposons: genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. *Annu. Rev. Biochem.*, **55**, 631–661.

Wetterstrand, K. (2012) DNA Sequencing Costs: Data from the NHGRI Large-Scale Genome Sequencing Program. `http://www.genome.gov/sequencingcosts`. Online; accessed 2012.05.17.

Wicker, T., Stein, N., Albar, L., Feuillet, C., Schlagenhauf, E. and Keller, B. (2001) Analysis of a contiguous 211 kb sequence in diploid wheat (*Triticum monococcum* L.) reveals multiple mechanisms of genome evolution. *Plant J.*, **26**, 307–316.

Wicker, T., Buchmann, J.P. and Keller, B. (2010) Patching gaps in plant genomes results in gene movement and erosion of colinearity. *Genome Res.*, **20**, 1229–1237.

Wicker, T., Guyot, R., Yahiaoui, N. and Keller, B. (2003a) CACTA transposons in Triticeae. A diverse family of high-copy repetitive elements. *Plant Physiol.*, **132**, 52–63.

Wicker, T. and Keller, B. (2007) Genome-wide comparative analysis of *copia* retrotransposons in Triticeae, rice, and *Arabidopsis* reveals conserved ancient evolutionary lineages and distinct dynamics of individual *copia* families. *Genome Res.*, **17**, 1072–1081.

Wicker, T., Krattinger, S.G., Lagudah, E.S., Komatsuda, T., Pourkheirandish, M., Matsumoto, T., Cloutier, S., Reiser, L., Kanamori, H., Sato, K. *et al.* (2009a) Analysis of intraspecies diversity in wheat and barley genomes identifies breakpoints of ancient haplotypes and provides insight into the structure of diploid and hexaploid triticeae gene pools. *Plant Physiol.*, **149**, 258–270.

Wicker, T., Matthewsand, D. and Beat, K. (2002) TREP: a database for Triticeae repetitive elements. *Trends Plant Sci.*, **7**, 561–562.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O. *et al.* (2007a) A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.*, **8**, 973–982.

Wicker, T., Taudien, S., Houben, A., Keller, B., Graner, A., Platzer, M. and Stein, N. (2009b) A whole-genome snapshot of 454 sequences exposes the composition of the barley genome and provides evidence for parallel evolution of genome size in wheat and barley. *Plant J.*, **59**, 712–722.

Wicker, T., Yahiaoui, N., Guyot, R., Schlagenhauf, E., Liu, Z.D., Dubcovsky, J. and Keller, B. (2003b) Rapid genome divergence at orthologous low molecular weight glutenin loci of the A and A^m genomes of wheat. *Plant Cell*, **15**, 1186–1197.

Wicker, T., Yahiaoui, N. and Keller, B. (2007b) Contrasting rates of evolution in *Pm3* loci from three wheat species and rice. *Genetics*, **177**, 1207–1216.

Wicker, T., Zimmermann, W., Perovic, D., Paterson, A.H., Ganal, M., Graner, A. and Stein, N. (2005) A detailed look at 7 million years of genome evolution in a 439 kb contiguous sequence at the barley *Hv-eIF4E* locus: recombination, rearrangements and repeats. *Plant J.*, **41**, 184–194.

Witte, C.P., Le, Q.H., Bureau, T. and Kumar, A. (2001) Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes. *Proc. Natl Acad. Sci. USA*, **98**, 13778–13783.

Wolfe, K.H., Gouy, M., Yang, Y.W., Sharp, P.M. and Li, W.H. (1989) Date of the monocot-dicot divergence estimated from chloroplast DNA sequence data. *Proc. Natl Acad. Sci. USA*, **86**, 6201–6205.

Wolfgruber, T.K., Sharma, A., Schneider, K.L., Albert, P.S., Koo, D.H., Shi, J., Gao, Z., Han, F., Lee, H., Xu, R. *et al.* (2009) Maize centromere structure and evolution: sequence analysis of centromeres 2 and 5 reveals dynamic loci shaped primarily by retrotransposons. *PLoS Genet.*, **5**, e1000743.

Wu, L. and Hickson, I.D. (2003) The Bloom's syndrome helicase suppresses crossing over during homologous recombination. *Nature*, **426**, 870–874.

Xu, M., Brar, H.K., Grosic, S., Palmer, R.G. and Bhattacharyya, M.K. (2010) Excision of an active CACTA-like transposable element from *DFR2* causes variegated flowers in soybean [*Glycine max* (L.) Merr.]. *Genetics*, **184**, 53–63.

Yang, G., Nagel, D.H., Feschotte, C., Hancock, C.N. and Wessler, S.R. (2009) Tuned for transposition: molecular determinants underlying the hyperactivity of a *Stowaway* MITE. *Science*, **325**, 1391–1394.

Yang, G., Weil, C.F. and Wessler, S.R. (2006) A rice *Tc1/mariner*–Like Element Transposes in Yeast. *Plant Cell*, **18**, 2469–2478.

Yeadon, P.J. and Catcheside, D.E. (1995) *Guest*: a 98 bp inverted repeat transposable element in *Neurospora crassa*. *Mol. Gen. Genet.*, **247**, 105–109.

Yu, G.L., Bradley, J.D., Attardi, L.D. and Blackburn, E.H. (1990) *In vivo* alteration of telomere sequences and senescence caused by mutated *Tetrahymena* telomerase RNAs. *Nature*, **344**, 126–132.

Yu, J., Hu, S., Wang, J., Wong, G.K.S., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X. *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. indica). *Science*, **296**, 79–92.

Yuan, Y.W. and Wessler, S.R. (2011) The catalytic domain of all eukaryotic cut-and-paste transposase superfamilies. *Proc. Natl Acad. Sci. USA*, **108**, 7884–7889.

Zabala, G. and Vodkin, L. (2007) Novel exon combinations generated by alternative splicing of gene fragments mobilized by a CACTA transposon in *Glycine max*. *BMC Plant Biol.*, **7**, 38.

Acronyms

AP aspartatic proteinase

BAC bacterial artificial chromosome

bp base pair

CDS coding sequence

CNS conserved non-coding sequence

CRM centromeric retrotransposons of maize

CRR centromeric retrotransposons of rice

DSB double-strand break

EDT estimated divergence time

IR illegitimate recombination

INT integrase

kb kilo base pair

LARD large retrotransposon derivative

LTR Long Terminal Repeat

Mbp mega base pair

MITE Miniature Inverted Repeat Transposable Element

ML maximum-likelihood

MP maximum-parsymony

Myr million year

MYA million year ago

NGS next generation sequencing technologies

ORF Open Reading Frame

PEC paired-end complex

RFLP restriction fragment length polymorphism

RT reverse transcriptase

siRNA small interfering RNA

SSA simple sequence annealing

SDSA sequence dependent strand annealing

TE transposable element

TIR Terminal Inverted Repeat

TRIM terminal repeat in miniature

TSD target site duplication

UECO unequal crossing over

WGS whole-genome shotgun

Acknowledgments

The last years were full of changes and discoveries, on a professional and private level. I would like to use this opportunity to thank and acknowledge the people who accompanied me on smooth and rough rides across the PhD universe, turning the last years in an extraordinary experience.

I want to express my gratitude to PD Dr. Thomas Wicker for excellent supervision and Prof. Dr. Beat Keller for the opportunity of doing my thesis in his group and funding in the last months. I enjoyed a lot of freedom and experienced a creative atmosphere. Both demonstrated great patience and interest in my projects and taught me the tricks of the trade. I appreciate their good advises (usually to think twice) and the always fast and thorough corrections. I also thank Christian von Mering for being part of my PhD Committee and Prof. Dr. Alan H. Schulman to be my external reviewer. The greater part of this thesis has been funded by the Swiss National Foundation (SNF) grant 31003A-122242.

I want to thank the people in the Keller group as they created a friendly atmosphere in the lab and were always good company. I could rely on their help in scientific and wetlab questions (happened sometimes), especially Gabriele Büsing and Gerhard Herren. The help of Beatrice Senger and Liselotte Selter during my short excursion in planting *Brachypodium* (also this happened) is not forgotten. For pointing out weak points in my arguments, good advises and realistic descriptions of the near future I thank Prof. Dr. Robert Dudler. For late lab sessions and Tahini imports I'm grateful to Dr. Roi Ben David.

Special thanks and a hug go to the inhabitants of the P3-12 Lab:
Thomas Wicker, Simone Oberh11100100nsli, Margarita Shatalina, Konstantinos Kritsas, Stefan Roffler, Christopher Middleton and the virtual member, Claire Wicker. It is hard to find better companions for such a journey and beyond.

I want to acknowledge the hundreds of volunteers in the GNU and Linux community who write and maintain the software which made my own tools running.

I have to thank my friends outside the lab for their interests in my doing, their stamina waiting for me being late for appointments and ongoing friendship. The time together, either spent at concerts, outside or where ever did a fair share in surviving this PhD: Dennis Slajh, Mario Isler, Gabriel Neurohr, Thomas Ammann, Nadine and James Breen, Tim Kamber and Simone Peter.

I thank Peter Vollmar for his early support and apologize for the further delay of a book of mine.

A special gratitude goes to Gabriele Büsing for her love, patience and continuous support of my work. Without her I would have hardly stuck my head out of the ivory tower, giving me the needed distance to actual problems which often led to new ideas. When needed, she also got my feet back on the ground.

Last but not least I am deeply grateful to my family. Especially my parents, Anna and Hans Buchmann and my brother Lukas. Their love, support and interest through all those years since the beginning were crucial and kept me going. They always had an open ear and door, offering a save harbor in troubled times.

Curriculum vitae

Surname:	Buchmann
Name:	Jan Piotr
Date of birth:	16.06.1982
Place of origin:	Zürich ZH

Education:

1998–2002:	Academic Upper Secondary School
	Gymnasium Freie Katholische Schulen Zürich (FKSZ), Zürich
2002:	Matura in modern linguistic studies
	Major in Italian language
2002–2007:	Undergraduate studies in Biology, University of Zurich
2007–2008:	MSc thesis in the group of Prof. Dr. Beat Keller
	Supervised by Dr. Thomas Wicker
	Title: Comparative genomics of orthologous sequences between *Brachypodium sylvaticum*, *Brachypodium distachyon*, rice and sorghum derived from the *Sr2* resistance locus in wheat
February 2008:	MSc in Biology, Plant Sciences, University of Zurich, Zurich
March 2008 – present:	PhD thesis in Prof. Dr. Beat Keller's group under the supervision of Dr. Thomas Wicker

Appendix A

This appendix contains the additional tables for Chapter 2. Table A.1 describes the identified exon/intron boundaries in the analyzed *CACTA* transposases. Table A.2 is the summary of all compared exon/intron boundaries.

Table A.1 Exon/intron boundaries and the corresponding coordinates for *CACTA* transposases on the protein sequence. The boundaries are numbered im the 5′ to 3′ orientation.

	boundary			
CACTA transposase	1	2	3	4
Korbin	510	720	814	854
Baron	523	732	826	865
Chester	526	734	829	865
Sherman	832	959		
Storm	828	956		
CACTA_I	835	956		
Preston	743	848		
CACTA_K	745	851		
Sandro	749	847		
CACTA_F	847			
CACTA_H	839			
Janus	838			
Joey	844			
CACTA_J	496	751		
En1	884			
Norman	889			
DOPPIA	844	881		
Baldur	732	840		
Seamus	731	836	875	
Balduin	850	888		
CACTA_G	848			
Alfred	887			
Isaac	861	902		
Isidor	859	896		
Rufus	852	889		
Radon	842	877		
Dario	727	843	902	
Aron	858	906	1020	
Horace	713	892		

Table A.2 Comparison of intron/exon boundaries between the investigated *CACTA* transposases. Numbers 1 to 4 found right or below *CACTA* name indicate the exon/intron boundary on the corresponding transposase in 5′ to 3′ direction. The coordinates for the correspondong boundary can be found in (Table A.1). Values in the table indicate the signal confidence: ●: High confidence (++); ◆: low confidence (+); ■: putative conserved (?+); –: not conserved (−);

		Korbin				Baron				Chester				Sherman		Storm		CACTA_I		Preston		CACTA_K		Sandro		CACTA_F	CACTA_H	Janus	Joey	CACTA_J		En1	Norman	DOPPIA		Baldur		Seamus			Balduin		CACTA_G	Alfred	Isaac		Isidor		Rufus		Radon		Dario			Aron		
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	1	2	1	2	1	2	1	2	1	2	1	1	1	1	1	2	1	1	1	2	1	2	1	2	3	1	2	1	1	1	2	1	2	1	2	1	2	1	2	3	1	2	3
Baron	1	●	–	–	–
	2	–	●	–	–
	3	–	–	●	–
	4	–	–	–	■
Chester	1	●	–	–	–	●	–	–	–
	2	–	●	–	–	–	●	–	–
	3	–	–	●	–	–	–	●	–
	4	–	–	–	■	–	–	–	●
Sherman	1	–	●	–	–	–	●	–	–	–	●	–	–
	2	–	–	–	◆	–	–	–	◆	–	–	–	◆
Storm	1	–	●	–	–	–	●	–	–	–	●	–	–	●	–
	2	–	–	–	◆	–	–	–	◆	–	–	–	◆	–	●
CACTA_I	1	–	■	–	–	–	■	–	–	–	■	–	–	■	–	■	–
	2	–	–	■	–	–	–	–	◆	–	–	■	–	–	–	–	–
Preston	1	–	●	–	–	–	●	–	–	–	●	–	–	●	–	●	–	■	–
	2	–	–	■	–	–	–	–	■	–	–	–	■	–	■	–	–	–	–
CACTA_K	1	–	●	–	–	–	●	–	–	–	●	–	–	●	–	●	–	■	–	●	–
	2	–	–	–	■	–	–	–	■	–	–	–	■	–	■	–	■	–	–	–	■
Sandro	1	–	●	–	–	–	●	–	–	–	●	–	–	●	–	●	–	■	–	●	–	●	–
	2	–	–	–	■	–	–	–	■	–	–	–	■	–	–	–	–	–	–	–	■	–	◆
CACTA_F	1	–	–	–	■	–	–	–	■	–	–	–	■	–	■	–	■	–	–	–	■	–	■	–	■
CACTA_H	1	–	–	–	■	–	–	–	■	–	–	–	■	–	■	–	■	–	–	–	■	–	■	–	■	◆
Janus	1	–	–	–	■	–	–	–	■	–	–	–	■	–	■	–	■	–	–	–	■	–	◆	–	◆	■	■
Joey	1	–	–	–	■	–	–	–	■	–	–	–	■	–	■	–	■	–	–	–	■	–	◆	–	◆	■	■	●
CACTA_J	1	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
	2	–	–	–	■	–	–	–	■	–	–	–	■	–	■	–	■	–	■	–	■	–	◆	–	◆	◆	◆	◆	◆
En1	1	–	–	–	◆	–	–	–	◆	–	–	–	◆	–	■	–	■	–	–	–	■	–	◆	–	◆	–	–	–	–	–	■
Norman	1	–	–	–	◆	–	–	–	◆	–	–	–	◆	–	◆	–	◆	–	–	–	■	–	■	–	◆	■	■	■	■	–	■	●
DOPPIA	1	–	–	◆	–	–	–	◆	–	–	–	◆	–	–	■	–	■	–	■	–	–	–	■	–	■	■	■	■	■	–	■	–	–
	2	–	–	–	■	–	–	–	■	–	–	–	◆	–	–	–	–	–	–	–	■	–	–	–	–	–	–	–	–	–	–	●	◆
Baldur	1	–	●	–	–	–	●	–	–	–	●	–	–	–	–	–	–	–	–	●	–	–	–	–	–	■	■	■	■	–	–	–	–	◆	–
	2	–	–	◆	–	–	–	■	–	–	–	◆	–	–	◆	–	◆	–	■	–	■	–	■	–	■	–	–	–	–	–	■	◆	◆	–	■
Seamus	1	–	●	–	–	–	●	–	–	–	●	–	–	●	–	●	–	■	–	●	–	●	–	●	–	–	–	–	–	–	–	–	–	–	–	●	–
	2	–	–	◆	–	–	–	◆	–	–	–	◆	–	–	–	–	–	–	–	–	–	–	–	–	–	■	■	■	■	–	–	–	–	◆	–	–	◆
	3	–	–	–	◆	–	–	–	◆	–	–	–	◆	–	◆	–	◆	–	■	–	–	–	■	–	◆	–	–	–	–	–	■	◆	◆	–	◆	–	–
Balduin	1	–	–	◆	–	–	–	◆	–	–	–	◆	–	–	–	–	–	–	–	–	–	–	–	–	–	■	■	■	■	–	–	–	–	◆	–	–	◆	–	◆	–
	2	–	–	–	◆	–	–	–	◆	–	–	–	◆	–	◆	–	◆	–	◆	–	■	–	■	–	■	–	–	–	–	–	■	◆	◆	–	■	–	–	–	–	–
CACTA_G	1	–	–	■	–	–	–	■	–	–	–	■	–	–	◆	–	◆	–	■	–	■	–	◆	–	◆	■	■	◆	■	–	■	■	■	■	–	–	■	–	–	■	–	■
Alfred	1	–	–	–	◆	–	–	–	◆	–	–	–	◆	–	◆	–	◆	–	–	–	■	–	■	–	◆	–	–	–	–	–	–	◆	◆	–	◆	–	–	–	–	◆	–	◆	■
Isaac	1	–	–	◆	–	–	–	◆	–	–	–	◆	–	–	–	–	–	–	–	–	■	–	–	–	–	■	■	■	■	–	–	–	–	●	–	–	■	–	◆	–	◆	–	■	–
	2	–	–	–	◆	–	–	–	◆	–	–	–	◆	–	◆	–	◆	–	■	–	–	–	■	–	◆	–	–	–	–	–	■	◆	◆	–	◆	–	–	–	–	◆	–	◆	–	●
Isidor	1	–	–	◆	–	–	–	◆	–	–	–	◆	–	–	–	–	–	–	–	–	–	–	■	–	■	■	■	■	■	–	–	–	–	●	–	–	–	●	–	–	–	–	■	–	●	–
	2	–	–	–	◆	–	–	–	◆	–	–	–	◆	–	◆	–	◆	–	■	–	■	–	–	–	–	–	–	–	–	–	■	◆	◆	–	◆	■	–	–	–	◆	■	–	–	◆	–	●
Rufus	1	–	–	◆	–	–	–	■	–	–	–	◆	–	–	–	–	–	–	–	●	–	–	–	–	–	■	■	■	◆	–	–	–	–	●	–	●	–	●	–	–	◆	–	■	–	●	–	●	–
	2	–	–	–	◆	–	–	–	◆	–	–	–	◆	–	◆	–	◆	–	■	–	■	–	■	–	■	–	–	–	–	–	■	◆	◆	–	●	–	–	–	–	◆	–	◆	–	◆	–	●	–	●
Radon	1	–	–	◆	–	–	–	◆	–	–	–	◆	–	–	–	–	–	–	–	–	■	–	–	–	–	■	■	■	◆	–	■	–	–	●	–	●	–	●	–	–	●	–	■	–	●	–	●	–	●	–
	2	–	–	–	■	–	–	–	■	–	–	–	◆	–	◆	–	◆	–	■	–	–	–	■	–	◆	–	–	–	–	–	–	◆	■	–	■	–	–	–	–	◆	–	●	–	◆	–	●	–	●	–	●
Dario	1	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	■	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
	2	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
	3	–	–	■	–	–	–	◆	–	–	–	■	–	–	■	–	■	–	■	–	–	–	■	–	■	■	■	■	■	–	■	■	■	–	◆	–	■	–	–	■	–	■	◆	◆	–	◆	■	–	■	–	■	–
Aron	1	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	■	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
	2	–	–	■	–	–	–	◆	–	–	–	■	–	–	◆	–	◆	–	■	–	■	–	◆	–	■	■	■	◆	◆	–	■	◆	◆	–	◆	–	■	–	■	–	–	■	◆	■	–	■	–	■	–	■	–	■	●	–	–
	3	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	■	■	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	■
Horace	1	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	●	–	●	–	–	–
	2	–	–	■	–	–	–	■	–	–	–	■	–	–	–	–	–	–	–	–	■	–	◆	–	■	■	■	■	◆	–	■	■	■	–	–	–	■	–	■	–	–	■	–	–	–	■	–	–	–	■	–	–	–	–	●	–	■	–